A STATISTICAL INVESTIGATION OF PROCEDURE IN HOSPITAL CLINICS

S. D. WALTER

DEPARTMENT OF STATISTICS, UNIVERSITY OF EDINBURGH

Submitted for the degree of Doctor of Philosophy in the University of Edinburgh

1972

Declaration

The following record of research work is submitted as a Thesis for the Degree of Doctor of Philosphy in the University of Edinburgh, having been submitted for no other Degree. The work was carried out under the supervision of Professor D. J. Finney, F.R.S., of the Department of Statistics, Edinburgh University, and also with the help of Professor Eric Samuel of the Department of Medical Radiology in the Royal Infirmary, Edinburgh. The Thesis has been composed by me, and except where due acknowledgement is made, the work is original.

Stephen D. Walter

Stephen D. Walter

June, 1972

Abstract of Thesis

This thesis is a study of the daily working routine of a hospital X-Ray department, and it describes the empirical variation observed in a real department by the use of mathematical models; these models may be used to predict the effects of changes in departmental operating policy, thereby leading to a more effective use of the available resources.

An outline is made of the working procedures of a typical X-Ray department, and there is also a survey of the literature. The various types of variability to be observed are described, together with some efficiency measures which may be adopted. In Chapter 4 the results of survey work in the Royal Infirmary, Edinburgh are presented in the form of an analysis of the work-load on the X-Ray department, with respect to its constitution, origin and distribution over the X-Ray facilities. An investigation is made into the relationship between the time to perform a given examination, and the age, sex and mobility of the patient. It was found that considerable differences exist between the service time distributions of patients with different characteristics, and these were later used to consider if improvements in efficiency might result by dealing with patients in homogeneous groups. A number of comments are made on particular problem areas in the Royal Infirmary.

Chapter 5 deals with mathematical queueing theory models of an X-Ray department. A use of a method of Smith is demonstrated to obtain the Laplace transform of the queueing time distribution of the system $E_{k_1}/E_{k_2}/1$; the transform is a function of the roots of a certain transcendental equation.

The model is extended to include repeated examinations caused by technical failures. When an assumption of a constant probability of success at each stage is made, the transcendental equation is of a rather simpler form, and the queueing time distribution may be expressed as a weighted sum of exponential variables. By letting k_1 become large whilst keeping constant the service time mean, results are derived for the system $D/E_k/1$; again the transcendental equation is somewhat simpler than in the general model.

The above models have an input either of units arriving in batches at regular intervals (to represent patients with appointments), or of single units arriving at random (non-appointment patients). To investigate a system where, as in reality, these two classes of patient are mixed in some ratio (r), the model denoted by $(M + D_m)/M/1$ is considered. The easy generalisation to $(M + D_m)/E_k/1$ is demonstrated later. The inter-arrival intervals in this system are not independent, and it is shown that an assumption of independence is unrealistic. The algebraic steady-state solution is derived for the distribution $\underline{\pi}$ of the number of people in the queueing system just after the appointment times; the solution involves an infinite set of equations which are approximated to obtain numerical results. Approximations to the equation coefficients are also used, and an analysis of both the resulting errors is made.

Moments and quantiles of the queueing time distribution for the regular arrivals are given numerically. For random arrivals, these are given algebraically, together with bounds for the moments. The results for $r = 0$ and $r = \infty$ are shown as numerical solutions of the systems M/M/1

and D/M/1. The time-dependent solution of $(M + D_m)/M/1$ is derived and used to see how rapidly the moments of $\underline{\pi}$ converge to their equilibrium values when the initial distribution is given. The results indicate that an assumption of equilibrium behaviour may be justified in many cases.

Chapter 6 describes some computer simulations of X-Ray departments working with a traffic intensity of one. Values of the average doctor idle-time per clinic and the average patient queueing time are estimated, and comparisons made between clinics with different lengths, values of the ratio r, service time distribution, batch size and initial number. For given maximum levels of idle time and queueing time, it is possible to select values of the other parameters which give clinics of the greatest efficiency. It is found that improvements always result when the proportion of appointment patients is increased, and sometimes there are substantial gains. Clinics dealing mainly with patients from certain sources are also simulated. Because of the differences in the distributions of age and service time of patients from different sources, it is found that even by a simple grouping of patients into in- and outpatient sets, reductions are possible in both the doctor idle time and the queueing time for all patients.

Although this work is based on the study of X-Ray departments, the methods used in it are sufficiently general that they could be applied to other types of clinic with only slight modification.

Table of Contents

Page

Page

List of Figures

Page

List of Tables

Acknowledgements

I should like to thank Professor D. J. Finney, F.R.S. for his constant guidance and encouragement during this research. I am also very grateful to Professor Eric Samuel, of the Department of Medical Radiology in the Royal Infirmary, Edinburgh, whose practical assistance in many ways in the hospital was invaluable. The friendly co-operation of all the staff of this department during the survey work in the Royal Infirmary was appreciated. I would also like to thank those individuals, both in the hospital and in the Statistics Department of Edinburgh University, who took a particular interest in the work and made a number of useful suggestions. Finally I would like to thank my typists, particularly Mrs. G. Hamilton who carried out her task so efficiently despite the distance between her and myself.

1. Introduction

Since the discovery of X-Rays only seventy-five years ago, and the subsequent development of their use in radiodiagnosis, many branches of medicine have undergone a remarkable transition. Doctors today recommend investigations which would have been inconceivable a few years ago. The introduction of new isotopes, improved precision in engineering and the development of techniques in the use of machines are leading to the perfection and wide use of new types of examination. There is now mass X-Ray screening, using relatively inexpensive film; in general the reduction in the cost and effort involved in producing finished plates has resulted in an increased number of exposures being regarded as routine for many examinations. Some investigations requiring long sequences of observations have now become commonplace. In short, the X-Ray department is now a recognised and thriving component of the modern health service.

The function of the X-Ray department within its hospital is highly specialised, and has unique organisational problems as a result. With the recent increase in the volume and complexity of the demands being made on this part of the health services, the vital balances required for a smooth running organisation are not always maintained. The taking of X-Ray exposures may now be a familiar and accepted thing amongst the general public, but few relish the delays caused by long queues of patients who may on occasions wait several hours for the completion of their examinations. At other times radiographers, radiologists, technical staff and machinery may all stand idle for lack of work. These two situations may even arise at the same time in different areas of the department. It is possible that this

condition may even disturb the delicate relationship between radiologist and patient, which is necessary for a difficult examination, and this may defeat the progress which would otherwise have been possible.

It is clear that despite any personal sympathy one may have with the doctor or patient, we must adopt a rational approach in our procedures in order to allow the art of applying radiodiagnostic techniques to keep pace with the rapid development of technical achievement in this field.

The aim of this research was to examine the role of the X-Ray department in a modern hospital, construct models of its structure and working, and to use these models to suggest alternative administrative procedures which would result in a service which was, in some sense, more efficient than at present. The two major objectives of this work were firstly to determine how the existing facilities in a given department with a known demand for its services can be most effectively used to satisfy that demand, and secondly to estimate what facilities and resources are needed in order to satisfy a given demand with maximum effectiveness. The first problem is the more basic, as the second requires a solution of it to determine the "effectiveness" for any given system.

It was beyond the scope of this work to provide an absolute directive on every aspect of working procedure, but a rather more general approach has been adopted, and in addition several specific aspects of administrative policy have been investigated. The problems of this study originated in a radiodiagnostic department, and it was here that all the practical work was

carried out. However the theoretical queueing work of Chapter 5, and the simulations of Chapter 6 are developed in some generality, and with only minor modifications it would be possible to use the same approach and methodology in other types of hospital clinics.

1.1 Operational Research in the Health Services

The role of the radiodiagnostic department in a modern hospital is unique. It is commonly one of the largest departments providing a specialised service to the rest of the hospital, and as its daily routine is largely concerned with movements of people, it is set apart from some other service departments, such as Clinical Chemistry, which deal mainly with specimens. Here we have the first of many administrative restrictions placed on an X-Ray department, in that we may not have as much freedom in committing our patients to a queueing process, as if our incoming units were inanimate. Apart from the treatment of its patients, the X-Ray department must also provide a full report of the examinations, often immediately, to be used by the requesting agent. This element sets it aside from all therapeutic and curative departments which keep their own records and may have to report to no-one else. X-Ray must maintain its own specialised medical and technical staff, and considering the wide areas of application of the subject, we see that this complex of people, materials, and daily routine must put the X-Ray department in a unique category.

It is rare to find a department which deals with almost all the other units of the health service in the area, namely the wards and clinics of its own and of other hospitals, doctors, and members of the public directly. It has to provide a service which is convenient and

effective for its users, whilst simultaneously balancing the cost of running the department and giving itself a practical and efficient working procedure.

The application of operational research techniques to health service situations is comparatively new, even within the brief existence of operational research as a whole. Many of the original problems in this subject arose in industry, where one is usually attempting to maximise some function of the system, a measure of production quality and profits for example, subject to certain limiting constraints. Even where queueing theory is used in industry it is often possible to quantify a loss function when items have to queue to be processed, or a production line stands idle. We may then adjust the parameters under our control to, say, minimise our expected loss under our operating strategy.

In a public service context such as this, the only analogy which may be made with profit is the rapid completion of patients' treatment. The quality of the treatment may be affected by the operating system: for example there is evidence to suggest that doctors will work faster in order to clear a long queue of patients. Although a faster working rate by the doctor may not necessarily imply poorer service to patients, it is clearly undesirable to induce the doctor to work faster than his natural rate, either consciously or subconsciously, by placing him in a difficult working environment. The "profit" and quality here affect the customer so fundamentally that we must be prepared to lose some efficiency, in the industrial sense, for the provision of a more

acceptable patient procedure. This question of balancing two objectives, minimising cost and improving service, is fundamental to administrative procedure in this context.

Again in contrast to an industrial problem, there is great difficulty in quantifying losses arising in the queueing process. The problem of comparing the relative values of times wasted by doctor and patient is one that has only recently been considered at all. Not many years ago, appointment systems in general practice and out-patient clinics were rare. Uncompromisingly, the doctor's time was considered infinitely more valuable than that of his patients, and it was thought that the doctor should stand as small a chance as possible of having to be idle in waiting for patients to arrive. It was the norm, therefore, that all patients arrived at 8.30 a.m., and the doctor at 9 a.m. to begin work with a full waiting room. Only in recent years has it been widely recognised that a balance can and should be struck between the two waiting times, and appointment schemes are in fact now used almost universally in general practice, and fairly widely in hospital clinics.

However, in an X-Ray department, this problem is only one of many which must be considered by the administration when setting up appointment booking schemes. In a general practice surgery, almost all of the patient "service" is carried out by the doctor himself, including many peripheral activities such as prescription writing. From the patient's point of view, the routine is simple: he must queue only once to see the doctor, and is free to leave immediately or very shortly after the end of the consultation. In X-Ray, in addition to balancing patient and

doctor waiting times, the system must also have an overall effectiveness in its whole operation, that of providing with reasonable promptness a report of the examinations performed. As we shall see later in greater detail, the patient's progress through the department is much more complex than in a general practitioner's surgery; he may have several different types of examination requiring separate machinery, and therefore may have to queue several times. He may even have to queue to be examined several times on the same machine if a sequence of exposures at intervals is required, or if a previous exposure was of inferior quality. Further, there are patients who require their completed X-Rays and report before leaving the department, and these must also wait a considerable time while the film is processed, the diagnosis made and the report produced.

Another factor is the great expense and maintenance costs of the X-Ray machinery, which further complicates the waiting time issue. It would seem that something better is needed than the glib economic answer that patient waiting-time should be increased (with a corresponding reduction of machine idle time) for those patients using more expensive machines. In comparison with other medical departments, X-Ray presents great difficulties in the diversity of sources of its patients who arrive requiring a wide variety of attention, including some emergency treatment.

Compared with an industrial queueing situation, many of the factors affecting efficiency are fixed, or outside the control of the department. For example, the demand pattern on a given department will be reasonably

constant from within the hospital, and much of the demand will be dictated by the size, nature, and working methods of other parts of the hospital; an afternoon visiting time or ward-patients' resting time may exclude certain times of the week for clinics. Good diplomatic relations with the other units of the hospital are obviously desirable to achieve compromises in matters of this sort, but the X-Ray administration will not, and should not expect to be able to operate in an environment where all scheduling decisions are to its own advantage.

We must note that the rationale of longer established operational research in industry, where many conflicting objectives may more easily be put on a common economic basis, are not applicable here. Ethical questions are prominent, and the scientifically-minded operational researcher must greatly respect the subjective judgement of medical staff, which may be built on years of practical daily experience. Health administrators are usually constrained to work within narrow ranges of the parameters under their direct or indirect control, and the subjective element will be to the fore when choosing from a limited number of compromise procedures.

We have seen that within the current rise of interest in the application of scientific methodology to health service administration, the particular operational procedure of X-Ray departments presents new and challenging problems to be investigated.

1.2 General Description of X-Ray Department Procedure

Figure 1.1 shows most of the principles governing procedure in a typical X-Ray department, by considering the path of a patient through it, and some of the activities which concern him. The figure may be used in conjunction with sections 1.2.1 to 1.2.4 which deal with some of these aspects in turn.

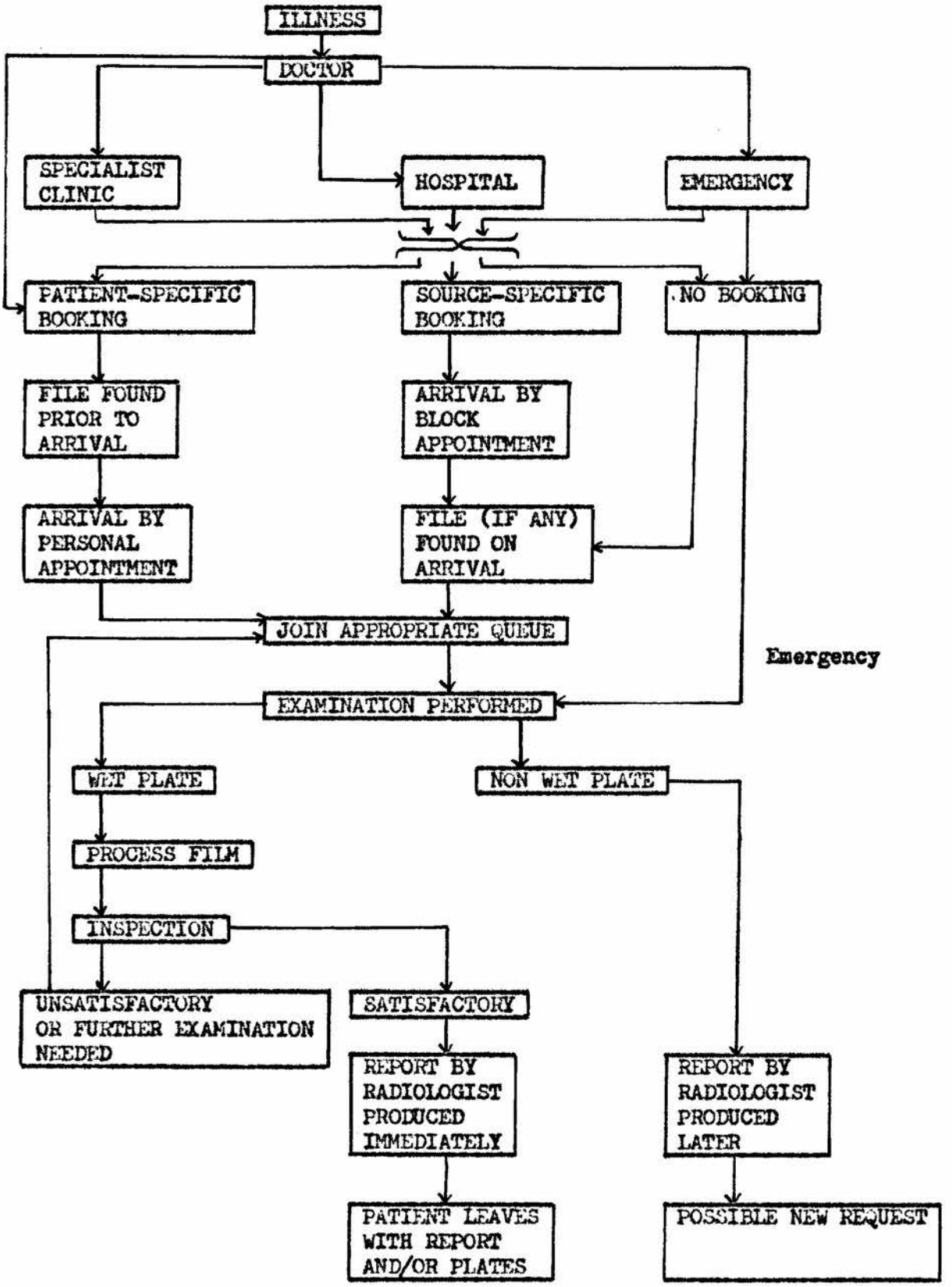
1.2.1 Origins of Patients

Patients arrive from many different sources, but these can be grouped into four main classes. The relative numbers from each source vary between departments, and will also depend on local practice. To maintain generality, no attempt has been made here to rank the importance of these sources. Firstly, there are patients who consult their own doctor, who then sends them directly for X-Ray. Secondly there are patients from specialist clinics within the department's own hospital, and also ward patients. Thirdly in large hospitals or teaching hospitals, patients may also be sent from other smaller hospitals lacking extensive X-Ray facilities themselves. Fourthly there is the emergency category made up of accident patients requiring immediate or fairly urgent attention, and who arrive without prior warning to the department.

1.2.2 Patient Arrival and Reception

On arrival at the department, patients may be put into one of three groups, depending on how much prior warning the department has had of the patient's arrival. These groups are:-

FIGURE 1.1 REPRESENTATION OF SOME PATIENT-RELATED ACTIVITIES IN A TYPICAL X-RAY DEPARTMENT



1. Patient-specific appointments. Included in this category are patients for whom a particular time in a clinic has been allotted. This is commonly the case for complex examinations requiring much doctor or machine time, or sometimes if the patient has difficulty in attending, for example having a long distance to travel. Also in this class, by common practice, are patients referred for examination by local general practice doctors or other hospitals, but the incidence of this will depend on local attitudes towards this procedure. There are, too, patients who require examination at a particular time for medical reasons, pre- or post-operatively, say; some patients for whom special difficulty is expected are scheduled at off-peak times for the department's own convenience.

2. Source-specific appointments. If a clinic regularly sends a number of patients for an X-Ray examination of a particular type during its working hours, sometimes time is reserved in the X-Ray department for the partial or total use of one room by those patients. For example, an Ear Nose and Throat clinic block booking may be made for the sinus machine on one or two mornings a week. Less often, a hospital ward may decide to book a session for, say, a routine chest examination of a group of its patients. In practice it is rare to find a clinic which will yield a sufficient number of patients to justify complete use of one of the rooms in the main X-Ray department, but in a hospital of reasonable size, there are often a number of specialities which have a special X-Ray unit in their own clinics.

3. No appointment. These are patients who arrive without any prior warning to the department. This may happen when a person is sent from a clinic which has no regular arrangement with X-Ray, from an outside doctor or by a ward doctor. Also included in this class are the emergency patients. If a hospital is not sufficiently large, it may not have a separate X-Ray unit in the casualty department, and in this case emergency patients have to be dealt with using the same facilities as the other people. Patients in a critical condition and requiring immediate attention will of course receive first priority throughout the system, to the inconvenience of other people who will have to wait longer as a result. In hospitals where there is a complete casualty department with its own X-Ray unit this problem does not arise as a rule. Only when the examination cannot be performed on the rather limited equipment generally available in Casualty does it become necessary to disturb the schedule of the main department. Even with a separate casualty department, there may still be queue-jumping by the acutely-ill patients. For the purposes of work scheduling by the X-Ray administration, these random arrivals become more important at times when there are several clinics open and therefore likely to be sending patients. The distribution of numbers of patients arriving from each of these sources may be known from past experience, and thus be somewhat predictable.

On the receipt of a prior request for an X-Ray, clerical staff will, whenever possible, locate the patient's previous records, if any, and they will be held in readiness on the day

of the appointment. Consideration of the information from the previous files and exposed plates may have an influence on the new examination to be performed, as it is the radiologist who finally decides which examinations shall be performed, and not the requesting doctor. For an unscheduled arrival, except in the rather unusual event of a person bringing his own file (from another hospital, say), this clerical operation must be carried out on the patient's arrival, and the patient can make little progress through the department until this has been done.

1.2.3 Queueing and Examination

The patient now proceeds to the appropriate area of the department where he either waits for his appointment or joins a general queue for the facilities needed for his examination. Commonly, the queue discipline is "first come, first served", and this is administratively the most simple. Depending on local conditions, some priority may be given to appointment patients, but it is unusual to see strict adherence to an appointment schedule, except, perhaps, in clinics performing longer examinations where the vast majority of patients are by appointment anyway.

Patients are classified in one further way before examination. Before they leave the department, some require a complete set of processed films and the radiologist's diagnostic report. This is usually for the use of a clinic doctor, or some other doctor who wishes to use the information derived from the X-Rays fairly rapidly. A patient may return to his clinic to continue

his treatment in the same session, or the same day, or a diagnosis may be needed for a ward patient before his treatment can continue or enter a new phase.

1.2.4 Film Processing, Diagnostic Reporting and Patient Departure

On a modern machine, processing the exposed films may take as little as ninety seconds, but may often take five minutes or more where less advanced equipment is in use. Once processed, the plates are then passed to a team of reporting radiologists who inspect them and dictate a diagnosis to clerical staff directly; the report and films are then presented to the patient, who leaves. Because of the comparative urgency of this method, the films for this class of patient are often known as "Wet Plates", and this term, in common X-Ray parlance, often describes the patient himself. A Wet Plate patient thus has to spend an additional period in the X-Ray department after the completion of his examination.

When it is expected that meaningful results may be difficult to obtain, any patient may be required to wait until his pictures have been processed and a radiologist, often the one who performed the examination, has made a brief inspection of them to assess their technical worth, but not necessarily until he has made his full report. For patients in both this and the Wet Plate category, an unsatisfactory exposure necessitates a re-examination, and the patient will frequently have to queue and wait through the whole cycle again.

Non Wet Plate patients, for whom no special technical difficulty is anticipated, may leave immediately on completion of the examination. The films are processed in the same way, but some time may elapse before it arrives at a reporting bench, and the diagnostic report is usually made onto a dictation machine to be typed later. The completed report is finally delivered by messenger. An exposure discovered to be of little use subsequent to the patient's departure may result in the recall of the patient for re-examination. Further, secondary examinations may also be requested for any patient on the basis of the previous investigation.

All reports and films are eventually returned to the files in the X-Ray department.

2. Survey of the Literature

Writers on the administration of hospital clinics have adopted a wide range of approaches. Articles by practising medical administrators, or others who deal with specific hospitals or problems, are often highly empirical; on the other hand, theoretical statisticians are continuing to develop the mathematics of queueing theory, which at present can examine only much simplified models of general situations. Most authors, however, wishing to maintain at least some generality in the applicability of their conclusions, and finding completely theoretical models of their system too inaccurate for their purposes, tend to adopt a method somewhere between these two standpoints. It is convenient at this point to postpone the study of those papers dealing solely with theoretical queueing work until Chapter 5, which deals with this aspect of the work. Here we will examine only articles which deal at least in part with the practical aspects of clinic administration, and queueing theory studies which arose with the particular application of the hospital in mind.

2.1 The Introduction of Appointment Schemes

It is only in relatively recent years that appointment systems have been used at all widely in the health service. After the war, radiography experienced a particularly rapid increase in the number of patients being examined, and appointment systems were introduced almost by necessity, and not always without their initial difficulties. Ashworth (1954) wrote:

> "In the initial stages of working by appointment, it was our custom to book cases, regardless of type, at rigid fifteen minute intervals. The only attempt at grouping different types of examination was in the case of barium work: of these four to six were booked consecutively from 9 a.m., again at quarter-hourly intervals. Also

> one or two short periods were reserved each week for screening chests and hearts. In spite of its many patent flaws, this scheme worked well enough in 1948 or 1949, but as you all know only too well, by that time departments were facing up to a rapid expansion of the scope and volume of work, often without increase in accommodation or equipment, and sometimes without additional staff. Our rather rigid system, therefore, had to be drastically modified to meet this mounting pressure. It was clear, for example, that while it might take an average radiographer fifteen or twenty minutes to X-ray a lumbar spine, a dozen chests could be disposed of in the same time. Moreover, ... once apparatus is arranged for some specific type of examination, economy of time and effort results if a number of similar cases are radiographed consecutively."

These remarks illustrate the need to reduce inhomogeneity in the patient input, and this was achieved here by grouping patients by examination types. Many of the points noted now appear obvious to anyone with a knowledge of elementary queueing theory results, but at that time trial and error methods were still being used to improve efficiency. Ashworth observes that total grouping of patients was not possible because of lack of equipment, inconvenience to patients, and scheduling difficulties such as ward rounds and visiting hours. His excellent discussion of practical difficulties facing hospital administrators makes interesting reading, and much of the argument holds true today.

2.2 Simulation Models of a Clinic

In the early 1950's, Bailey and Welch produced an important series of papers when studying many aspects of the health service from a statistical viewpoint for the first time; in particular, outpatient and general practice clinics were under attention. Bailey (1952a) observed that the traffic intensity, the ratio of the rate of arrivals to the service rate, was close to unity in such situations, and he therefore did not feel justified in using the equilibrium queueing theory results then available. Roughly speaking, the closer the traffic intensity is to unity the longer

a queueing system will take to "settle down" and reach a steady-state or equilibrium condition. Thus in a clinic, where the intensity is almost one and the queueing process is usually of short duration, equilibrium theory may be an inadequate description.

Bailey's object was to find the relationship between wasted doctor time in waiting for the next patient to arrive (idle-time), and patient waiting or queueing time. Bearing in mind that the waiting time distribution seemed to be J-shaped with a long "tail", he considered percentiles of the distribution rather than basing his thesis solely on verbal arguments about average waiting times. His method was to simulate clinics of 25 patients each, the calculations being done by hand. His patients "arrived" at regular intervals, and he fitted a Pearson III, or gamma-type distribution to the service times; the service time distribution was based on data from a survey of outpatient clinics by the Nuffield Provincial Hospitals Trust (1955). The effects of the variance parameter of the gamma distribution, and the starting time of the doctor were investigated. A compromise between the expected idle time and waiting time was obtained when the doctor started work with the arrival of the second patient, but it was noted that the behaviour of the system depended critically on the appointment interval, and it was necessary to estimate the mean consultation time very accurately. Bailey's conclusions were preceded by some subjective arguments as to what levels of idle and waiting times would be reasonable.

Welch and Bailey (1952) remarked on the over-insurance of many doctors against a high idle-time. It had been observed that patients commonly waited in clinics for over an hour on average for a relatively short

consultation of perhaps only a few minutes: the patients arrived early as a rule, and the doctor often late. These authors also realised the advantages of segregating patients into homogeneous groups, this time into old and new patients:-

> "If old patients had one fixed consultation time, and new patients had a different fixed consultation time, then some waiting would be inevitable unless the two types of patient were kept separate, each having their own clinic with an appointment interval equal to the appropriate consultation time. Similar considerations apply when there is a statistical distribution of consultation times. If the average consultation times for old and new patients are different, as is usual, then it is more efficient to design appointment systems for the two types of patients. As a general principle, any class of patient with an average consultation time which is appreciably different from that of other classes should be dealt with on its own. If there are too few patients for a whole clinic, then the first part of the clinic might be confined to new patients who would be called forward at a rate appropriate to their average consultation time, while the second part of the clinic would be confined to old patients with a different appointment interval. In many hospitals this is already done."

Bailey (1955) noted that in a finite realisation of a queueing system, such as a clinic, the patients at the end of the queue would have a greater expected wait than those dealt with at the start. He refined his earlier simulation results by having three patients present at the start, and calling at a rather slower rate than previously, with a traffic intensity of less than unity.

In later papers in this series, Bailey makes a more global investigation into the clinic, which is seen as a unit within the hospital framework. He investigated the number of clinic sessions needed to deal with a given demand, regarding the problem as one of queueing theory with a list of patients who wait for a number of days before service is given, in bulk,

when a session is arranged. Later he formulated methods for assessing the demands on hospitals and their clinics by catchment areas.

2.3 Practical Studies

In the late 1950's Flagle and others published a number of articles concerning methodology in outpatient clinics in the United States. These are interesting for comparison purposes, but unfortunately many of the findings are inapplicable to a British clinic because of basic differences in the clinic environments. For example, when dealing with delays to patients, the American researchers found that one of the greatest bottlenecks was commonly at the accounting department, which is, of course, non-existent in a National Health Service clinic in Britain.

In 1959, the shortage of radiographers at that time and the expectation of a hospital building plan (H.M.S.O., 1962) prompted the first comprehensive study of the diagnostic X-Ray department. This was carried out by the Nuffield Provincial Hospitals Trust, and the results published as a book in 1962, "Towards a Clearer View". The investigating team visited six hospitals of different types throughout Britain, and studied each in depth. Although much of the work is concerned with internal aspects of efficiency, such as the use of radiologists' time and the versatility of equipment, stress was also laid on the relation between X-Ray and the other units of the hospital which the team felt had been treated only superficially by previous authors. "Towards a Clearer View" contains a list of principles evolved from that study, and recommendations for their implementation; the list is extracted in Appendix 1; some of these general topics are pursued further in this thesis. The Nuffield report goes into detail about many

specific aspects of X-Ray departmental routine, and reference to some of these also is made later in this study.

It may be seen from the literature of this time that the growing awareness of the need for efficiency and economy in clinics continued, both from the patient and staff points of view. The Office of Health Economics (1963) reported a 48% rise in the number of inpatients in British hospitals during the period 1949 to 1961, and an increase of investment in radiographic equipment from £18.2 million in 1953 to £22.5 million in 1961. On the patients' side, Rossiter and Reynolds (1963) cited many advantages of reducing patient waiting, including an annual saving of 2.6 million man-hours if waiting were halved. The hospital would also benefit by having to provide less waiting space, and a better doctor/patient relationship would be enjoyed.

Many researchers were trying to quantify "efficiency" in terms relevant to the health service situation. The Office of Health Economics (1967) itemised several factors, including the percentages of correct diagnoses and effective treatment, organisational efficiency, and the costs of inpatient and outpatient care. Rossiter and Reynolds suggest a definition of efficiency of $2c/t \times 100\%$, where c is the average consultation time and t is the average time spent by a patient in the clinic in total : t/c is thus the average number of consultation times spent in the clinic. This efficiency is 100% when one patient, on average, is waiting in line for attention, corresponding to when the mean queueing time is equal to one mean service time. A similar definition is adopted by Hardie (1955) which has three patients waiting, on average, for 100% efficiency. Once again subjective argument is needed to determine "reasonable" levels of waiting.

Efficiency measures of this type have obvious disadvantages. For a fixed value of c, it is possible in principle to increase efficiency (as defined by Rossiter and Reynolds) up to almost 200% by reducing to almost zero the queueing time of each patient; this would be done by providing a very large number of doctors and other staff, all working for only a small proportion of their available time. This sort of measure only involves waiting time and idle time as factors affecting efficiency; to be used sensibly, we must regard 100% efficiency as optimal, to take account of idle time increasing when the efficiency measure is greater than 100%.

Increasing awareness that appointment schedules must do more than insure against excessive idle time is reflected in a paper by Welch given as part of a colloquium on appointment systems in hospitals and general practice (Jackson, Welch and Fry, 1964). Welch stresses the need for the doctor to arrive punctually, pointing out the disastrous consequences, in terms of increased waiting for all patients, which results from even marginal lateness on the doctor's part. Rationalised appointment systems had now been successfully used in some clinics for many years: Welch reported the case of a particular casualty department where the waiting space had been reduced by some 75%. Fry, a general practitioner, also reported a dramatic drop in patient waiting at his partnership clinic when appointments were introduced. It was possible in this example to deal with three-quarters of the patients by appointment, and still leave enough flexibility to cope under emergency or extreme conditions. Jackson, in the third paper of the colloquium, gave an instructive example of using a computer simulation to design an appointment scheme for general practice use.

Pike (1963 a) and Blanco White and Pike (1964) studied the outpatient clinic and casualty department situations with a rather more rigorous statistical approach than had been used previously. Pike developed a queueing system using a consultation time distribution of mixed type, where the particular distribution for a patient corresponded to his class of examination types. He then built a mathematical model of an outpatient clinic, and, restricting his attention to a single consultation time distribution, was able to derive analytically a number of results concerning the expected waiting of successive patients, and the expected idle time of the doctor. He developed Bailey's 1955 work on the equalisation of the expected waits by successive patients, and also considered the effects of patient unpunctuality and patients failing to attend the clinic. The results for models with unpunctual patients were derived by computer simulation. In the same work, Pike developed a mathematical model of a casualty department, including unpunctuality and many other important practical details; he also gave an example of the practical introduction of an appointment system, embodying a mixture of analytical rigour and commonsense.

A useful description is made by Scott and Gilmore (1966) of outpatient clinics in the Edinburgh hospitals. This article deals in detail with the relationship of general practitioners with the hospitals and the "open-access" system of patient referral.

The Nuffield Provincial Hospitals Trust (1965) carried out an extensive survey of outpatient waiting, partly as a follow-up study to their 1955 work; a summary of the conclusions and recommendations is given in Appendix 2. The Ministry of Health (1958) had suggested, as a rough guide

to acceptable levels of patient waiting, that 75% of patients should be seen within half an hour, and not more than 3% to 5% should have to wait more than an hour. The Nuffield survey showed that at only 11 out of the 60 hospitals investigated was outpatient waiting within these specified limits. Pike had also shown that with many appointment schedules used in practice, it was not theoretically possible to achieve these standards of service. From this, and other conclusions of the Nuffield report, it appears that still more work on the implementation of efficient work schedules in practice is needed.

A further example of simulation methods is given by Jeans et al (1972). A computer simulation model was built of the X-Ray department at the Bristol Royal Infirmary using the empirical results of surveys of service time distributions, arrival patterns and other parameters. It was possible to predict the effects of changes in the resources of the department, staff or machinery for example, and in the arrival rates of patients from various sources. Although such highly empirical models are usually valid for only fairly small ranges of the system parameters, this work illustrates another powerful use of simulation, that of periodic monitoring of the efficiency of an appointment system and other areas of departmental routine.

2.4 A Study of X-Ray Work by Fraser (1969)

Simulation techniques were also used in this work which was based on a particular example, the hospitals of the Reading area. Although the work is of value, some of its arguments make a number of fundamental and highly idealised assumptions, some of which are so far removed from present

practice that I have doubts about the practical applicability of some of the findings.

A large part of this work is devoted to the simulation of a ficticious department, under various sets of admittance "rules" for the wards and clinics. Some inpatients requiring examinations needing preparation were given appointments, but others were called directly from the wards as an appropriate machine became available. Outpatients were classified on arrival as to their examination, and for each type of examination there was a preference list of machines on which it could be performed. As an outpatient arrived, the list was scanned until a suitable machine was found, and he was then allocated to it. If no machine was available, the patient was added to the queue. The system could be modified by allowing inpatients to join a single mixed queue with outpatients, and by varying the proportion of appointments, and the number of clinics sending patients.

2.4.1 Limitations of Fraser's Model

Under present working conditions in most hospitals, it is not possible to "scan" a preference list of machines for availability in this way; it would be necessary to do this for each incoming patient, and also patients waiting in queue lines. To achieve this ideal at least one member of staff having a complete knowledge of the state of the queueing systems in all areas of the department at all times would be required, who could therefore direct arriving patients to an area of lesser congestion. Clearly, much wasted patient and doctor time would be saved with such a person present, but until the day when the National Health Service provides telecommunications equipment for each area of the department, it is impossible in practice for one

person to have anything but a very general picture of the department at any particular instant in time. Even if such information were available on, say, a small computer, to make such a scheme work would require the addition to the staff of a highly trained co-ordinator with a very agile mind. Also, Fraser's system of calling ward patients as machines become available would be trying for both the patients and ward staff, and other departments which might require the patient's attendance. Previous arrangements have to be made to transport the patients, over considerable distances in some large hospitals.

There are other inherent disadvantages in the Fraser system. There is evidence to show that staff working on a versatile X-Ray machine, who are presented with a stream of patients of mixed examination types, will not work as effectively as when they are dealing with patients of one kind only (cf. Ashworth). There is also the possibility that allotting a patient to a machine used principally for special examinations will result in patients immediately behind him in the queue, requiring this specialist treatment, having to wait much longer; had the first patient waited slightly longer himself for a more suitable machine, the combined waiting times of all patients would have been much less.

To sum up, the idea of a receptionist acting in a co-ordinating and planning role in the allocation of work as it arrives is basically a good one, and is in fact one of the recommendations of the "Towards a Clear View" team. However the running of a scheme relying completely on these methods to direct the work flow seems difficult to contemplate at present in the majority of Britain's hospitals; as we have seen, these methods may

present difficulties not obvious on first consideration. A good staff nurse in many existing departments may be seen to work along these lines anyway and often much congestion and delay is avoided in this manner. Fraser's assumptions lead to a model of a department which is rather divorced from present practice; instead, we will make some rather less restrictive assumptions about the system, particularly concerning the patient input mechanism. Also, in general, less empiricism is used in this work than Fraser's.

2.5 Conclusion

To some extent the very rapid increase in the number of patients receiving X-Ray treatment in the post-war years has eased. The present tendency is for work to increase in the form of longer and more intensive investigations ("Towards a Clearer View", p.2). Most departments now have some years' experience of running appointment schemes. However with the total estimated expenditure on health services in 1971-72 at £1803.7 million at 1969 prices (Office of Health Economics, 1970), with wages forming 64% of the total budget (Office of Health Economics, 1967), in the current economic climate no effort must be spared to further increase the efficiency of all aspects of the health service; however, care must be taken in the clinics not to destroy the balance between the two conflicting objectives, of reducing both idle and waiting time, which has finally been established in the last few years.

3. Outline of Problem Areas

3.1 Introduction

In this chapter some of the problems likely to arise in hospital clinics are described and classified, and an outline of the areas dealt with in this study is made. In the following chapter a detailed survey is made of a case hospital, the Royal Infirmary of Edinburgh, to establish the particular difficulties encountered there, and to collect data forming the basis of theoretical models and simulation studies described in later chapters.

From a global viewpoint considering both internal and external aspects of the department's procedure, the following four areas might be considered to cover most of the general problems facing a hospital administration when formulating operating policy:-

A) The definition of efficiency measures for any given facility of staff and machinery with known operating characteristics, and a known work demand.

B) An adequate description of the variability of the system.

C) For a given department, the allocation and scheduling of a given work demand to maximise efficiency.

D) The provision of new facilities to deal with a given demand at some specified level of efficiency. Questions of this type clearly entail knowledge of the major aspects of C).

A brief description is given below of each of these topics in turn.

3.2 The Definition of Efficiency Measures

A comprehensive definition of departmental efficiency should make allowance for the following aspects of the system:-

(i) The total running cost of the department.

(ii) The utilisation of rooms and machinery.

(iii) The utilisation of staff, particularly highly trained members.

(iv) The level of inconvenience in the service to patients, for example in the form of queueing or waiting times.

(v) The total number of patients being treated.

(vi) The quality of results.

(vii) The overall ease of operation in the system.

Many of these factors are interdependent, and optimising one will often improve others; for example, increasing the percentage occupancy of examination rooms will also tend to improve staff utilisation and increase the flow of patients; improvements of working conditions and the ease of production of results will often improve the quality of the end result. The only real conflicts between these factors in the common objective of increasing efficiency arise when patient considerations are involved; inconvenience to patients may often only be reduced by impairing the level of another factor, such as cost or staff utilisation for example.

Much of this work is concerned with measures of efficiency using aspects of factors, (ii), (iii), (iv) and (v). If we are considering the maximisation of efficiency within a given department which has a stable demand and budget, then the total cost and total number of patients may be assumed constant, and neglected in the maximisation. Also, for a given department the quality of results seems to be only indirectly affected by the rate of patient flow, possibly because of extra strain on the staff. Deteriorations of standards are regarded extremely seriously by the medical profession, and we shall assume a high constant quality of results over the range of work intensities generally encountered in real clinics. Ease of operation will be mentioned indirectly

through a number of aspects; this is a rather difficult variable to quantify for a whole department, but as we shall see, improvements will occur through the detailed study of some of the other efficiency factors.

3.3 Description of System Variability

Many of the problems of administrators in providing clinics services arise because of random or unexplained variation in aspects of the system. If all facets of the internal and external structure of a department were exactly predictable, it would be possible in principle to achieve a given standard of service at minimum cost or maximum efficiency. However, the major aspects of any real clinic usually do demonstrate substantial variability, and all the consequent administrative difficulties. Once we have a complete description of the system and its variability, we may then attempt to allocate work to the available resources in some optimal manner. In practice, however, complete knowledge of the behaviour of the system under all circumstances will not be available.

Variability in the system may be grouped into two classes, (1) variation of external influences on the department, mainly the volume and constitution of the work demand, and (2) variation within the department; each of these is described below.

3.3.1 Variation of Work Demand

In the general description of a department in Chapter 1, we saw how the patient input was made up of a mixture of streams from several sources; rarely will any of these streams be constant over any reasonable length of time considered either within or between days. By the nature of the work, some patients will always arrive without appointments, introducing a random element

into the clinic. The Nuffield team in "Towards a Clearer View" adopted a classification of patients into groups of "Wholly Controllable", "Predictable" and "Wholly Uncontrolled". The first category included patients arriving by appointment or others whose arrival could be specified by the department; the choice of appointment times is subject to some constraints, such as the hours of radiologists' sessions, or rest periods for ward patients, for example. "Predictable" consisted mainly of patients from outpatient departments whose distribution of demand was known from past experience. "Wholly Uncontrolled" was made up of casualty and emergency patients. Casualty cases may arrive at any hour, and may be first priority patients; otherwise this source of patients may be regarded in the same way as any other specialist department, as its demand pattern is also known from experience.

In this work a classification of patients into "appointment" and "random" groups is made; the definitions of these and the relationship to the "Towards a Clearer View" groups is made clear later. This classification is important, as later stages of this work attempt to quantify the effects of changing the demand pattern by policy decisions such as increasing the ratio of appointment to non-appointment cases, or dealing with the work load in sessions of patients from a restricted number of sources each. Effects of changing the total demand on the department are also considered.

3.3.2 Variation within Department

A major potential cause of congestion is the variation in service times of patients undergoing the same examination. A large number of factors may affect this statistic, including the following: the age, sex, and mobility of the patient; the machinery and staff performing the examination; the work intensity or patient flow through the facility; the length of the queue; the

number of patients examined previously during the session; the degree of heterogeneity in the patient stream - staff may work faster if presented with a sequence of similar cases; time of day. Previous workers have not investigated fully these aspects, despite the fact that such variation will almost invariably cause congestion and patient queueing. Such variation, if correlated with the origins of patients, might well lend weight to the idea of segregating patients by source into separate working sessions.

For reasons other than the above, the time to complete the processing and diagnostic report will be variable. Further indeterminacy arises in most of the other areas of the system, such as clerical and unskilled aspects, and machine breakdowns.

3.4 Optimal Allocation of Work to Available Resources

Even assuming that we have a complete knowledge of the behaviour of the system under all possible circumstances (never possible in reality), there still remains the separate problem of allocating the work to the facilities available in the department. Once again, our object is to maximise the efficiency of the system as indicated by some function of the factors itemised above. In this section we are concerned with an existing department with a known demand being made on it, and so factors (i) and (v), the total running costs and total number of patients, are taken as constant. The relation of the other factors to the allocation problem are described in turn below.

3.4.1 Utilisation of Resources and Staff

Various measures of these factors may be adopted. For material resources, the one often used is the percentage patient occupancy of a room compared to the total number of working hours available; an alternative is the total number

of patients examined in each room per day. The "Towards a Clearer View" team found the average room occupancy at its six sample hospitals to be between 30% and 50%. One of the report's recommendations is that full use should be made of trained staff such as radiologists and radiographers, with as much aid as possible to carry out unskilled and ancilliary tasks for them. One may measure the effectiveness of the use of staff by observing individuals at a set of random time points, and noting the task on which they are currently engaged - a so-called "activity" sample. This was done by the Nuffield team on radiographers, and it was found that only 20% to 40% of their time was spent in performing the tasks for which they had been specifically trained. In many hospitals, large proportions of their time were spent in carrying out ancilliary tasks such as clerical work and film processing. Walking about the department accounted for 8% to 23% of radiographers' time in different hospitals.

An investigation by the North West Metropolitan Regional Hospitals Board Organisation and Management and Work Study team confirmed these general findings: as an example of problem solving it was observed that by relieving radiographers of clerical work and spreading the work load throughout the day, their productivity rose by 20%. In the same study, similar increases were achieved in the productivity of dark-room technicians and clerical staff by fairly limited modifications to premises, machinery and procedure.

3.4.2 Patient Waiting

To the majority of patients, the length of queueing and waiting times to be endured is one of the most important factors in determining their satisfaction with the service. There may be other influences such as personal convenience in the time of day of the examination; however a high standard of professional practice is assumed by both patient and staff.

A good service for the patient may also profit the department by increased patient-staff co-operation, and overall ease of operating the service. Apart from these particular aspects, improvement of the service to the benefit of the patients usually implies a decrease in efficiency as measured by one of the other factors, often the total running costs. Thus concern for the patients' well-being is the main influence acting "against" the department's interests, in an economic sense. In recent years, though, it has been recognised that these patient considerations are of importance, and operating policies have been adjusted accordingly, for example by the introduction of appointment schemes.

Queueing time and waiting time are variables which are relatively easy to quantify and observe. They and the complementary doctor idle-time may be affected by many things, and in fact may be expected to change as a result of a procedural alteration in almost any area of the department. Any comprehensive study of waiting in a particular department should take account of the following:-

1) The co-operation with other departments and wards of the hospital in order to predict work demand patterns. These may be regulated to advantage by the judicious timing of appointments and provision of staff, and may be kept fairly constant by these means.

2) The policy adopted for assigning priorities to patient categories.

3) The provision of emergency services. The advantages and disadvantages of providing an integrated or separate casualty facility should be carefully considered.

4) Improvement of efficiency in darkroom and clerical functions.

5) The delegation of responsibility to staff.

6) The use of labour-saving devices, and the more general question of efficient room layouts.

7) The provision of waiting space and its layout.

8) Co-ordination of the activities in different rooms or areas of the department.

9) The maintenance of a co-operative patient-staff relationship, for example by informing patients of the causes of any delay.

Many of the above topics will naturally have a bearing on some of the other efficiency factors, for example the use of staff time.

3.4.3 Quality of Results

There are indications that the quality of both service and results is adversely affected when the system is put under high stress or working intensity.(For an example, see Pike (1963a)). Such tendencies are not taken lightly by the medical profession and steps must often be taken to avoid them. The medical staff may react to an increased work load by working rather longer hours and maintaining the same high standard of service. If, however, it becomes clear that the system is seriously overloaded, the remedy must be in terms of an enlargement to the department with extra staff or machinery. Suggestions for quantifications of this quality variable are the time devoted to each patient, or the percentage of correct diagnoses.

3.4.4 Ease of Operation

This is another variable rather difficult to quantify, substantially affecting the efficiency criteria adopted. It may be measured indirectly through some of the other variables considered above, and use may be made of the subjective judgement and experience of the staff to determine the overall effectiveness of the system, and establish ways of improving it.

3.5 Provision of New Facilities

This is not dealt with explicitly in this work, but, as previously noted, many of the methods and solutions to problems in the previous section may be applied here. The building of a new department also involves many additional problems, such as estimating the future demand by such methods as Bailey's catchment area process (Bailey, 1956). These particular problems are outside the scope of this work.

3.6 Outline of Problems considered in this Study

It was not the object of this study, nor was it possible to study each of the above in detail; many problem areas are treated indirectly during the work, and others are omitted entirely. It was rather more the intention to make a general investigation of some aspects, and not to become too involved with establishing details of efficient procedure for particular departments. Investigations such as the Nuffield team reports, while they may be excellent in their own right, tend to produce recommendations like "reduce the number of inpatient chest cases in the morning, and deal with them at a slacker time in the afternoon", or "stagger staff lunch breaks to allow the continuous use of rooms". It should be stressed that the general principles and recommendations laid down in "Towards a Clearer View" form a very good basis for the improvement of any particular department.

It did seem that a good theoretical description or model of a clinic queueing system was lacking; in particular little seemed to be known either quantitatively or qualitatively of how major sources of variation, such as differences of service time distribution with age, or parameters such as the proportion of appointment patients, affected the queueing behaviour. The

construction of a theoretical description of the real system constituted the first major portion of the work. Secondly, this description, useless in isolation, was used to establish further general policy outlines on a quantitative basis, and strengthen the existing methodology to be applied to individual departments. The efficiency factors used were mainly those affecting the patient queueing time and doctor idle time, which seemed to be at the crux of many organisational difficulties.

It is clear that there is a host of theoretical and practical problems in the field of clinic administration, and some of them may form the basis of future research. This thesis being primarily one of statistical research treats some problems in a manner which may at times seem rather detached from reality. We have seen, however, that there are advantages in such a general approach, avoiding local detail, and it is hoped that relevant practical outlines of policy may be formed on the basis of the results. The closing chapters of this study contain some suggestions of possible ways of increasing efficiency as indicated by the theory developed here.

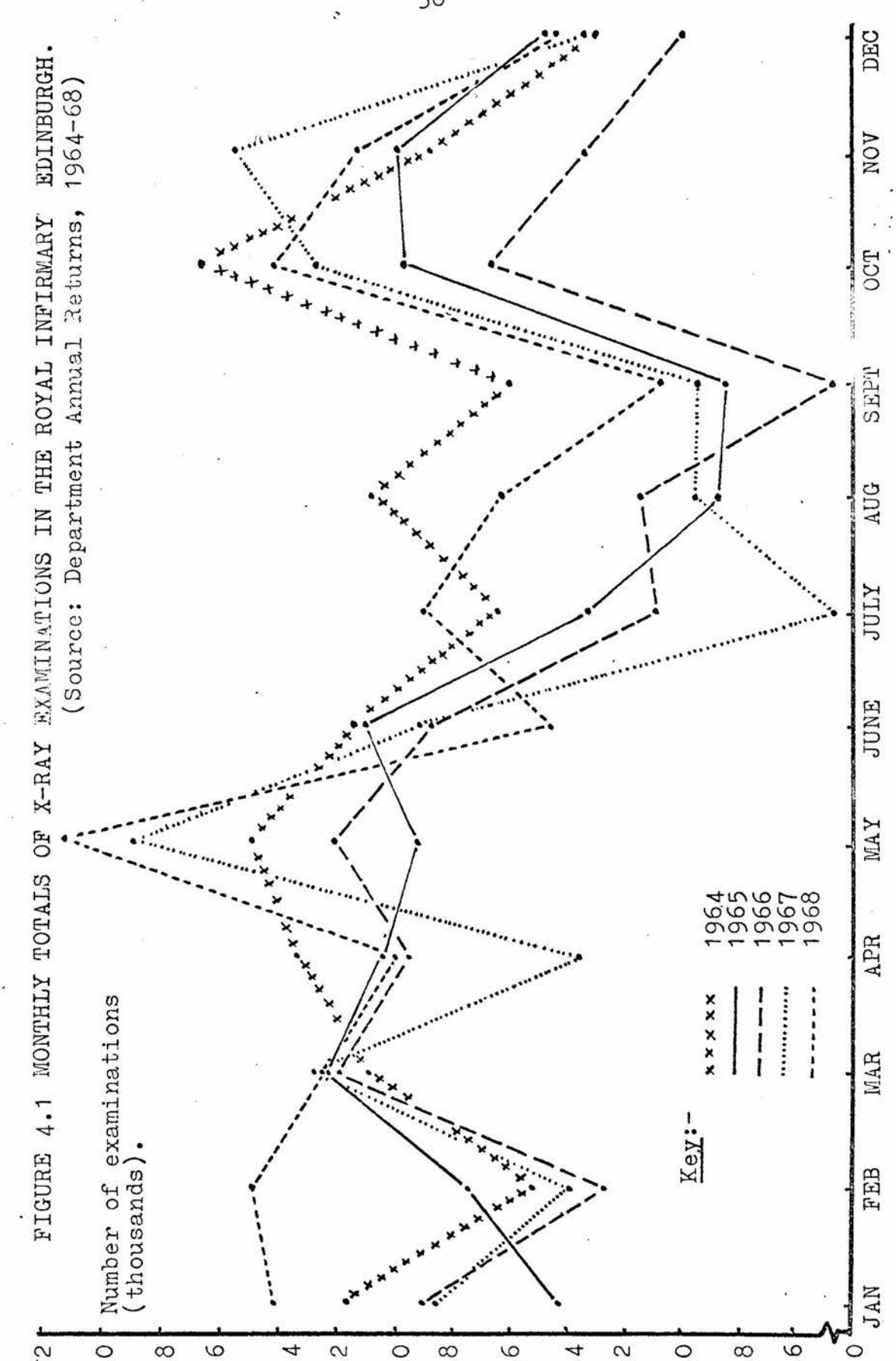
4. Survey of X-Ray Work in the Royal Infirmary, Edinburgh

4.1 Introduction

This chapter describes the practical work and data collection carried out in the Royal Infirmary of Edinburgh, and outlines the particular administrative difficulties observed there. The results and analyses from the data form the basis of the theoretical and simulation studies described in the following chapters. Although these methods are founded directly on observations made in only one hospital, it is hoped that they are sufficiently general to be applied in other types of departments and hospitals. Indeed there are indications from previous investigations that there is a remarkable consistency in the procedural methodology and behaviour of clinics and hospitals of apparently different types. The Nuffield report selected six hospitals of different types and found many of their problem areas to be similar, despite a wide variety in the type of work handled and the methods used. With this in mind, it was hoped that the difficulties observed in the Royal Infirmary would conform with those discussed in the principles and recommendations of "Towards a Clearer View".

The Royal Infirmary of Edinburgh is a large teaching hospital of some 1280 beds, and much of the clinical teaching within the medical school of Edinburgh University is carried out there. The total number of X-Ray examinations performed at all the hospital's facilities is running at approximately 120,000 annually. Figure 4.1 shows the monthly totals for 1964-68. We see that over these years there has been no substantial deviation from a stable monthly distribution which has the following salient points: lower totals in the summer months July to September; peaks in May, prominent in some years; a peak in either March or October.

FIGURE 4.1 MONTHLY TOTALS OF X-RAY EXAMINATIONS IN THE ROYAL INFIRMARY EDINBURGH.
(Source: Department Annual Returns, 1964-68)



Ideally a hospital administration would like to cater for a demand which is always constant in number and constitution. Highly trained medical staff cannot be employed on a seasonal basis, and for specialist paramedical workers such as radiographers there is no possibility of transfer to other tasks of equal responsibility. In general terms the administration must accept this seasonal variation in the work load and plan accordingly. There are however a considerable number of X-Ray examinations, particularly chests, which are performed routinely as a check-up; many of these may even originate from within the hospital, as in the case of staff examinations. Any attempt to schedule such examinations for less busy months can only result in better utilisation of staff and facilities during these slack periods, and will help to alleviate congestion at busier times of year.

The radiodiagnostic resources at the Royal Infirmary at the beginning of this study consisted of a general purpose Main department of ten major rooms, a two room Accident and Emergency department (or Casualty), and several smaller specialist departments in other parts of the hospital. Table 4.1 shows the distribution of work, in terms of examinations performed, between the departments of the hospital. Since the time that the data in this chapter were collected, there have been a few additions to the Accident and Emergency X-Ray facilities.

As in most large departments, the X-Ray work at the Royal Infirmary was performed in a number of rooms, each dealing usually with only a limited number of examination types. There were also contingencies to deal with machine breakdowns or emergencies, as indicated in the following list of rooms and the examinations normally performed in them.

TABLE 4.1 DISTRIBUTION OF X-RAY WORK IN THE ROYAL INFIRMARY OF EDINBURGH

Department	Percentage of total examinations
Main	56.4
Accident and Emergency	16.9
Orthopaedic	16.2
Obstetrics and Gynaecology	4.5
Neuro-radiology	3.7
Cardiology	1.9
Diagnostic Theatre	0.3

(Source: Department Annual Returns, 1964-68)

Room 1: (Screening room). Barium swallows, meals and enemas.

Room 2: All Barium work; some screening; urethrograms, sinograms; bronchograms; first stage of cholecystograms; special examinations.

Room 3: (Abdomen room with two machines). Intravenous pyelograms; last stage of cholecystograms; bone work; arteriograms if vascular room out of order; abdomens.

Room 4: (Spine room; two examination tables with one X-Ray tube) Spines; extremities; intravenous pyelograms (preferred to room 3).

Room 5: Skull work.

Room 6: (Preparation room with no machine.) Occasional injections given here.

Room 7: Tomography; sinuses; chests; intravenous pyelograms if spine room out of action.

Room 8: General radiology; screening; special swallows; first stage of cholecystogram; urethrogram.

Room 9: Chests.

Room 10: Planography.

Room 11: (Vascular room); arteriography.

Room 12: Mammography.

Rooms 10, 11 and 12 were used only for specialist sessions.

4.2 Daily Routine in the Main Department

To obtain a general picture of the behaviour of the department throughout the day and the operating procedures used, a week was spent in February 1970 in making some general observations on the amount of patient waiting, queue lengths, rates of working, and other aspects of the system.

All the staff were on duty by 9 a.m., but some rooms were slack or idle

until 10 a.m. because of a shortage of porters to collect inpatients from the wards. The darkroom usually began work continuously at about 9.30 a.m., and the chest room was seen to be particularly slow to receive a steady stream of cases. Inpatient work arrived in a steady stream until 11 or 11.30 a.m., and by 10.30 or 11 a.m. there was often a queue of patients because of an additional influx of outpatients and others from clinics. Outpatients had often to return to their specialist clinics with processed plates and completed diagnostic reports (Wet-plate patients), and so the rate of such patient arrivals decreased after 12 noon; patients arriving after this time would have insufficient time to be examined and return for further treatment at their clinic before the end of the session. A small number were treated over the lunch period, and these would return for an afternoon session at their clinic. The backlog of patients would usually be cleared by about 1 p.m.

After a slack lunch hour, the same pattern would be repeated over the afternoon working hours, with rather less patients in total. The peak period of coincidental arrivals from wards and clinics was less marked than in the morning, but still evident. Slack periods tended to be shorter at the start of the afternoon session than the morning one; it seemed to be easier for the porters to collect inpatients as they had fewer other tasks to do at this time, and there would often be a few patients dealt with during the lunch hour, thus tending to keep the department working. The peak period would often be delayed if ward patients had a rest period at about 2.30 p.m., or substantially reduced on Wednesday afternoons when there was a visiting hour.

Ward patients were likely to arrive until about 4.30 p.m., although the department was generally very quiet from 4 p.m. until the closing time at 5 p.m. Late arrivals would mean that the processing laboratory would work until 5.30 p.m.

or sometimes later, as it was the policy to clear each day's work before leaving.

In addition to the author's observations, extra information on patient queueing and the time to produce a diagnostic report was obtained with the assistance of the staff. During the initial period of observation in the department, the author attempted to acquaint as many staff members as possible with the purpose of the study. Notices, as in Figure 4.2, were posted at all appropriate points round the department. Then under direction, special forms were attached to a sample of patient request cards as they were dealt with in the reception area. The times that the card or patient reached certain points of the process were recorded by the appropriate staff member on the special timing form, a specimen of which may be seen in Figure 4.3 About 150 such forms were completed on a variety of patient types. Because of the fairly large number of people involved in making the observations, and other reasons, the data derived from this sample was not considered of a quality high enough to warrant any detailed analysis. However they were useful in drawing rather general conclusions about areas of congestion in the department, and some of the rough figures quoted later are based on this sample.

Patient queueing times in the first part of the morning were of the order of 10 minutes for all examinations; in the 10.30 - 12.00 rush period, this increased to an average of 20 to 30 minutes for simple examinations, and as much as 40 to 60 minutes for others. Also the variation increased substantially during the peak period. Patient queueing at comparable times in the afternoon was rather less, because of the reduced number of cases being dealt with in total.

FIGURE 4.2 INFORMATION SHEET ISSUED DURING THE SURVEY OF DEPARTMENTAL WORKING

TIMING SURVEY

These notes are intended to assist you in the completion of the special forms attached to certain patient cards. As you may be aware, it is not the purpose of this survey to "check up" on anyone's efficiency in his or her job. It is part of an independent investigation of operational methods in X-ray departments. Its success depends largely on the accuracy of your observations, and this may lead eventually to an improvement in the work management in the Department. On the timing forms please note the following:

1. NUMBER. The ordinary patient record number.

2. SOURCE. Fill in a ward number, clinic etc. as on the ordinary card.

3. EXAMINATION PERFORMED. For multiple or repeat examinations please record the type at the head of each new column.

4. TRANSPORT. Tick "chair", "walking", or "trolley" as appropriate.

5. W.P. Tick if Wet Plate.

6. APPOINTMENT. Leave blank if none.

Please fill in times 7 to 10 as the events occur, and don't be tempted to estimate them later. Use the hospital clocks whenever possible, as using watches may lead to inconsistencies.

7. RECEPTION. Arrival in the department.

8. START EXAMINATION. The time at which the patient is fetched from the queue or waiting area to begin his service. IMPORTANT. This is the time that the staff is ready, and actually about to start the examination.

9. FINISH EXAMINATION. Time when, depending on the examination type, the patient is merely waiting for a preliminary report on his plate(s), or the facilities are freed for the next patient. In a complex examination, where there are several stages involved, the time at the end of the last examination is required.

FIGURE 4.2 (continued)

10. RELEASE OR REPEAT. The time when the patient is released or when another examination is requested.

11. REPORT. For non wet-plate patients, the time the report is typed. For wet-plates, this is not necessary.

12. DELIVERY. For non wet-plate patients, the time the report reaches the department or ward which made the request, taken by the porter.

Times 8,9, and 10 may be repeated for multiple or repeat examinations. Fill in the corresponding times in successive columns.

FIGURE 4.3 SPECIMEN TIMING FORM

Walking	Chair	NUMBER			
Trolley		SOURCE			
Wet Plate	Sex	AGE			
		TIMES			
APPOINTMENT					
RECEPTION					
EXAMINATION PERFORMED					
START EXAMINATION					
FINISH EXAMINATION					
RELEASE OR REPEAT					
REPORT (with date)					
DELIVERY (with date)					

4.2.1 Time to Produce Diagnostic Report

Radiologists dictating diagnoses onto a machine (for non Wet-Plate cases) would often finish a recording tape in the late afternoon or early morning. Typists would be in greater demand for direct dictation for Wet Plate diagnoses during the two daily peak periods, especially the morning one, and machine tapes had secondary priority for being typed up. Once the patient request card had been typed with the diagnosis, checked and signed by the reporting radiologist, and matched with the exposed plates, they would be left for delivery by the department porters to the appropriate destinations. Often porters could collect or deliver a patient at the same time, but there were special trips daily specifically to deliver cards if necessary. If a case was unlucky at all the stages of processing, diagnostic reporting, typing and delivery, the turn-round could be as long as four days, but was usually one or two days.

At the beginning of this study in October 1969, the main department had a machine capable of processing and drying an exposed plate in seven minutes. This was replaced in April 1970 by a machine which processed it in 3½ minutes. This did appear to slightly reduce the turn-round time for Wet Plate patients, particularly those needing intermediate diagnostic inspections during the examination, but there was no significant reduction in the overall congestion during the mid-morning rush. For comparison, the casualty and orthopaedic machines both took five minutes to process and dry an exposed plate.

4.2.2 Combination Examinations

It appeared that a large proportion of the patients spending an excessive amount of time in the department were those requiring combination examinations in more than one area of the department. A small study in

March 1971 followed the progress through the system of some fifty such patients during the course of a week. Usually they had to queue for examination at each area, with no allowance being made for their previous queueing time; exceptions did occur when a staff nurse directed the radiographers to give priority to a patient on his second or subsequent examinations, but these were unusual. It was the practice for an inpatient with a combination examination such as, for example, a chest and abdomen, or chest and barium, to have the appointment booked for the more major examination early in the morning. This was essential for barium examinations which involve starving the patient for a short period beforehand; however for other examinations, the inpatient's arrival at the second examination room (usually the chest room) would often coincide with the morning peak. The work in some of the rooms performing more complex work did not have such a pronounced peak as the chest and bone rooms, and less slack or idle time at the start of the session. It would appear that something might be gained by giving some patients their examinations in the reverse order in such cases; this would tend to reduce peaks at all the examination rooms, and better utilise the chest room during times which are otherwise slack or idle.

4.2.3 Note on Survey Observations

Any investigation of this nature, involving observations of people at their jobs, naturally generates a certain amount of suspicion. This was most noticeable amongst the unskilled and non-medical workers who viewed all "time and motion" men as somehow "checking up" on their personal efficiency. While under observation all workers appear ill at ease to a greater or lesser extent, and behave in a manner which is sometimes atypical. It was sometimes necessary to ask several people what was the "true" picture, or normal

procedure in given circumstances; often much explanation and reassurance was needed before workers would accept that the investigation was not directed towards individuals, and was being conducted independently of the hospital administration.

4.3 Analysis of the Work-Load on the Department

4.3.1 Sample of Patient Records from the Departmental Archives

The constitution of the work demand being made on the hospital was investigated by taking a sample of the patient record cards from the departmental archives. Any patient who has not previously been X-Rayed at the hospital will be given a number which he will (in theory) use on all subsequent visits. Records from all the X-Ray departments except Obstetrics and Gynaecology were kept in a central file, and so it was the practice to allot blocks of patient numbers from time to time to particular departments. In due course these blocks would be included in the main file in numerical order. The allocation of patients to numbers is thus not strictly random, but a random sample of patient numbers will lead to unbiased estimators of the population parameters under investigation.

In November 1969, when the sample was made, the current patient record numbers being issued were just under 100,000. A set of uniformly distributed integers on the interval $[0,100,000]$ was selected using a table of random numbers. For every random number N chosen, details were taken from the five patient records numbered N, $N + 1$, ... $N + 4$. This procedure does not yield a sample which is strictly random, but again unbiased estimators are obtained. This device of selecting blocks of cards was a time-saver in both the generation of the random numbers and the location of the corresponding patient records.

The sample covered examinations performed during the period 1965-69 and the details of all visits of selected patients during this time were recorded; this naturally led to a slightly greater number of more recent examinations in the sample, a desirable feature. The random numbers selected led to data being collected from 850 patient record cards, one record representing one patient visit. This gave (by chance) exactly 1000 distinct examinations; a patient having, for example, skull and chest X-Rays will yield two examinations in the sample. The following details were noted from each card:-

1) Patient number (for reference)
2) Date of Examination
3) List of examinations performed
4) Origin of patient
5) Age of patient
6) Mobility of patient. On most cards was recorded a symbol to denote if the patient was walking, in a wheelchair, or on a trolley
7) The total number of X-Ray plates exposed
8) Wet-Plate patients. The request card was usually overstamped in such cases, and this was recorded
9) Total time spent in department. Space was included on the card for a note of the arrival and departure times of patients. However this was completed in only a small proportion of cases, and few conclusions could be drawn from the small number of completed cards in the sample. Later observations of the daily work routine provided data on this aspect of the study
10) Any other special details

4.3.2 Results of Sample

In Table 4.2 are shown the total number of examinations of each type recorded in the sample. The grouping of examinations in this way is sometimes a little arbitrary: the equipment of most hospitals is somewhat adaptable, and there may be a choice of locations for a given examination (for example intravenous pyelograms at the Royal Infirmary). Secondly it may be the local practice that examinations of the same group are performed routinely in different rooms: for example at the Royal Infirmary, Barium swallows are often performed in the chest room, as this is a simple examination compared to other barium work, and is easily carried out on a standard chest X-Ray machine.

In very general terms, the Royal Infirmary most closely resembles hospitals D and E of the Nuffield report: D undertook a substantial amount of advanced and new examinations, and E dealt with a high proportion of chest cases. Over the years of the sample, a slight increase was noticed in the proportion of some advanced examinations, for example using new isotopes; examples of these were arteriograms and tomography showing increasing demand and complexity. However the overall pattern of the work was roughly stable over the years considered.

A very slight decrease was noted in the average number of plates exposed for routine examinations; this possibly reflects a higher quality in the taking and production of plates, rather than a systematic policy of "fewer pictures" being applied. Although only a small amount of information was available from the sample, there were also indications that the number of patients affected by machine breakdowns had been reduced.

TABLE 4.2 DISTRIBUTION OF EXAMINATION TYPES

Chest work			
Chest	379		
Neck	6		
Thoracic inlet	6		
Ribs	4		
Clavicle	2	Total	397
Extremities			
Single extremity	154		
Multiple extremity	11	Total	165
Skull work			
Skull	50		
Sinus	32		
Mandible	10		
Mastoids	7		
Facial bones	5		
Dental	3		
Nasal bones	2		
Sialogram	2		
Orbits	1		
Multiple type	16	Total	128
Gastrointestinal (Barium)			
Meal	40		
Enema	23		
Swallow	16		
Series	7	Total	86

FIGURE 4.2 (continued)

Spine work			
Cervical spine	22		
Lumbar spine	11		
Pelvis/hips/sacro-illiac joints	8		
Skeletal survey	5		
Shoulder	5		
Lumbo-sacral spine	2		
Dorsal spine	1		
Multiple type	26	Total	80
Gastrointestinal (Abdomen)			
Abdomen	46		
Cholecystogram	7		
Cholangiogram	5		
Biligrafin	1	Total	59
Renal			
Intravenous Pyelogram (IVP)	31		
Cystogram	2	Total	33
Others			
Screening	35		
Angiogram	5		
Tomogram	5		
Arteriogram	4		
Vascular survey	2		
Mammogram	1	Total	52
		Grand total	1000

Table 4.3 shows a cross-categorisation of the examinations with the origins of patients. Tables 4.4 and 4.5 show the same grouping for multiple or combination examinations, that is of patient visits where more than one type of examination was performed. 13.9% of all patients required examinations of two types, including 11.5% involving a chest X-Ray. 1.4% needed three types of examination.

As examples, a patient requiring "wrist and ankle" examinations would be included in the "multiple extremity" category, and a "wrist and skull" patient would be classed as a combination examination of the extremity and skull groups. The first example would not be included in the combination examinations, but the second would.

There were 16 combination examinations of three types in the sample. All involved a chest X-Ray, and the other two examinations followed roughly the pattern of Table 4.5. Most sources produced patients of whom between 7% and 15% required combination examinations. One exception was Medical Out-Patients, whose proportion was estimated as 44%.

Some noteable features of Tables 4.3, 4.4 and 4.5 are the diversity of the demands from the Wards, and Medical and Surgical Out-Patients, and the "expected" concentrations of demand from certain sources, for example skulls from Ear, Nose and Throat, Extremities from Casualty, Chests from Staff, etc.

4.4 Distribution of Service Times

For an adequate description of the queueing system under study it is necessary to obtain some data on the time needed to complete examinations. We may obtain an empirical service time distribution for a particular examination,

TABLE 4.3 TOTAL NUMBERS OF EXAMINATION TYPES FROM VARIOUS SOURCES

Examination Group / Source	Chest	Extrems	Skull	Barium	Spine	Abdos	Renals	Others	Total
Wards	226	9	13	42	19	43	14	28	394
Casualty	27	109	41	0	18	6	0	0	201
Medical Outpatients	44	3	4	14	10	3	9	5	92
Ear Nose and Throat Cl	5	0	39	5	4	0	0	0	53
Surgical Outpatients	11	32	7	16	13	5	7	0	91
Staff	27	0	6	1	1	0	0	0	35
Radiotherapy	13	1	2	0	6	0	0	0	22
Ward Outpatients Cls	7	2	0	4	2	1	2	0	18
Coronary Care Unit	14	0	0	0	0	0	0	0	14
Oral Surgery Dept.	1	0	12	0	0	0	0	0	13
Cardiology	2	0	0	0	1	0	0	9	12
Peripheral Vascular Cl	4	3	0	0	2	0	0	2	11
Rheumatic clinic	4	6	0	0	0	0	0	0	10
Therapeutics	0	0	0	3	0	1	0	5	9
Other hospitals	3	0	1	0	1	0	0	2	7
Others	9	0	3	1	3	0	1	1	18
Total	397	165	128	86	80	59	33	52	1000

Abbreviations: Extrems - extremities; Abdos - abdomens; Cl - clinic.

TABLE 4.4 TOTAL NUMBERS OF COMBINATION EXAMINATIONS OF TWO TYPES INCLUDING CHEST FROM VARIOUS SOURCES

Examination Group	Extrems	Skull	Barium	Spine	Abdos	Renals	Others	Total
Source								
Wards	2	6	4	5	23	1	0	41
Casualty	2	2	0	1	3	0	0	8
Medical Outpatients	1	3	7	3	1	5	1	21
Ear Nose and Throat Cl	0	4	0	1	0	0	0	5
Surgical Outpatients	0	0	2	0	0	2	0	4
Staff	0	1	0	0	0	0	0	1
Radiotherapy	1	0	0	3	0	0	0	4
Ward Outpatient Cls	2	0	1	0	0	0	0	3
Oral Surgery Dept.	0	1	0	0	0	0	0	1
Peripheral Vascular Cl	2	0	0	0	0	0	0	2
Other hospitals	0	1	0	0	0	0	0	1
Rheumatic clinic	4	0	0	0	0	0	0	4
Others	0	1	1	1	0	0	0	3
Total	14	19	15	14	27	8	1	98

TABLE 4.5 TOTAL NUMBERS OF COMBINATION EXAMINATIONS OF TWO TYPES NOT INCLUDING CHEST FROM VARIOUS SOURCES

Examination Groups	Extremities	Barium	Spine	Abdomens	Total
Extremities			Casualty 1 SOP 1 PVC 1	Casualty 1	4
Skull	Casualty 4 Wards 1 SOP 2		Casualty 1 ENT 1 SOP 1		10
Barium			SOP 1		1
Abdomens		Wards 1 SOP 1	Wards 1 SOP 1		4
Renal			Ward O.P.1		1

Abbreviations: ENT - Ear Nose and Throat clinic; SOP - Surgical Outpatients; PVC - Peripheral Vascular clinic.

which will be useful in simulations, and possibly we will be able to approximate this distribution with a theoretical one, which may lead to some more rigorous mathematical analysis.

4.4.1 Standardisation of Data

It appears from consideration of past work in this area that there are considerable difficulties in standardising data and results referring to empirical queueing processes of this kind. Problems arise in the classification of examinations, from procedural differences between hospitals, and differences in measuring techniques used by observing teams. Consider as an example the results quoted in "Towards a Clearer View" in which the same investigating team measured the mean completion time of 99 examinations at each of six hospitals. Some typical results are shown in Table 4.6 for common examinations which occurred a reasonable number of times in the sample at each hospital. Even for the commonest single examination, the chest X-Ray, we may note substantial variation between hospitals, the mean time varying from 1.33 minutes to 4.09 minutes. Similar differences are to be found in the times for nearly all examination categories, and the differences appear to be neither consistent nor systematic.

Similar differences are evident between the results of different investigators. As an example compare in Table 4.7 "Towards a Clearer View" results and those of a Ministry of Health team quoted by Fraser; the table gives the figures for a few example examinations, and again large proportional differences are to be seen in most categories of the complete set of data. In contrast, note as an example the Lumbar Spine group which has very close means in both investigations, but had wide variation between hospitals. Briefly, there seems to be no way of standardising such data between hospitals or observing teams.

TABLE 4.6 COMPARISON BETWEEN HOSPITALS OF SOME MEAN SERVICE TIMES (MINUTES)

	HOSPITAL					
EXAMINATION	A	B	C	D	E	F
Chest	1.33	4.09	3.40	2.86	2.99	2.83
Hip	10.00	12.02	12.09	7.29	14.23	13.83
Lumbar Spine	5.17	7.78	11.36	14.21	9.46	9.56
Tibia + Fibia	4.53	4.93	9.59	4.21	10.79	5.44
Barium Enema	9.94	26.74	23.69	13.02	26.93	14.30

(Source: "Towards a clearer view", Appendix A)

TABLE 4.7 COMPARISON BETWEEN SURVEY TEAMS OF SOME MEAN SERVICE TIMES (MINUTES)

	SURVEY TEAM	
EXAMINATION	Ministry of Health	Nuffield
Chest	2.9	5.4
Hip	10.6	9.7
Lumbar Spine	8.5	8.4
Tibia + Fibia	5.6	6.7
Barium Enema	18.7	19.7
Knee	4.9	7.4
Nasal Bones	4.2	10.1
Dorsal Spine	12.1	9.8

(Sources: "Towards a clearer view", Appendix A and Fraser, Table 1.3.1)

4.5 Review of Study Objectives

At this stage in the study there were two main possible directions in which to continue. Firstly a large amount of data could have been collected at the Royal Infirmary in a manner similar to that of "Towards a Clearer View", and analyses made on a wide set of examination groups. This would yield specific quantitative proposals for efficiency improvement in this particular department, which would probably be of little use elsewhere. The alternative was to collect a smaller amount of data from the Royal Infirmary, make rather more assumptions about aspects of the behaviour of the system not covered by them, and to proceed to a rather more theoretical investigation of queueing systems of this kind. Retaining rather more generality in this manner would hopefully yield improvements applicable to a wider range of departments.

Unfortunately there was not sufficient time available to pursue both possibilities; the first course would have entailed a prohibitive amount of time to collect the data by one observer. As described earlier, some data were collected by the medical and administrative staff of the department at the Royal Infirmary; however it was felt that the data were not of a high enough quality to use the same methods for an exhaustive study of that department, quite apart from the rather unreasonable amount of time involved.

The second alternative was the one finally adopted; it was felt this was more in the spirit of the investigation, primarily one of statistical research. A large scale survey of one department would have duplicated previous work to some extent, and would have been more appropriately carried out by a team of workers. It was still the intention that the study would remain relevant to the hospital clinic situation, but would set up a methodology for establishing the principles affecting queueing systems of this type, rather than providing the numerical details for a particular department.

4.6 Chest Examinations

With the objectives of the previous section in mind, it was decided to make a fairly extensive investigation of one class of examinations, and a number of more cursory investigations of others. For a number of reasons, the chest examination was selected for the main study; this is a reasonably simple examination of short duration. At the Royal Infirmary, Chest X-Rays constituted an estimated 38% of the total number of examinations performed, and one room was set aside for almost exclusive use in this area. The stream of patients to this room was present at all times except very late in the afternoon and occasionally late in the morning; the usual peaks at 10.30 a.m. and 2.30 p.m. were evident. Selection of the Chest examination had the advantage that it was carried out commonly on patients of all ages, origins and mobilities. In addition, with the constant flow of patients and a fairly short examination, it was possible to observe a fairly large number of patients within a reasonable time.

4.6.1 Observations of Service Times

Observations were made in one week of March 1970 of some 256 chest X-Ray examinations. It was possible for one observer to record all the examinations during a session if required. The times that the observations were made covered a selection of busy and slack periods of the day at different times. The details noted were taken from the request card accompanying the patient, and were the same as those taken in the sample from the hospital archives, with the exception of the total time spent in the department; in addition the time to complete the examination was recorded.

As discussed in a previous section, there are difficulties in producing compatible sets of service times in different departments; in the Edinburgh

department patients waited for Chest X-Rays in a corridor area adjacent to the examination room, and were fetched when required by radiographers. The Nuffield team laid stress on differences in waiting times between patients who had to undress and those who did not. In the Edinburgh chest room, some outpatients were already sufficiently undressed from previous examinations; the majority of the other patients were not undressed, although a staff nurse sometimes directed patients to do so before being called for examination.

For patients able to walk, the only adjustments to the machinery needed were routine ones such as the height of the plate and X-Ray beam, and intensity of the beam. Patients in wheelchairs unable to stand often had the unexposed plate placed behind them in the chair, instead of in the normal frame; this procedure required further adjustments to the machine alignments. Trolley patients could sometimes sit upright, and they could then be wheeled against the X-Ray machine to expose the plate in the normal manner. When this was not possible, the plate was placed on the trolley under the patient, and the machinery rotated to give a vertical beam.

Chest X-Rays carried out for routine check-ups, often on hospital staff, were performed on a special machine, an "Odelca", using miniature film. For the purposes of fitting a theoretical distribution to the observed service times, only those examinations on the conventional machine were used. This was to ensure that all the observations came from the same underlying distribution, and also because the chest examination was the only one where such an alternative was used or required.

The time recorded for each examination was from when the radiographer called the patient until he had left the examination room. As more than one

radiographer was usually working in this room, there were sometimes overlaps in these intervals for successive patients, but usually of rather short duration.

4.6.2 Results and Fitting of Theoretical Distribution

Table 4.8 shows the observed distribution of service times for the chest X-Ray examination. The mean time was 2.72 minutes. The distribution is unimodal and has positive skewness. It was decided to try to fit a gamma, or Erlangian, distribution to the data.

We denote the service time by u, and assume a gamma distribution of mean b and "shape" parameter k. Then the probability density function of u is

$$f(u) = \frac{1}{\Gamma(k)} \left(\frac{k}{b}\right)^k u^{k-1} e^{-ku/b}$$

If the sample is of size N, and the observations are $u_1, u_2, \ldots u_N$, then the maximum likelihood estimates $\hat{b}$ and $\hat{k}$ of b and k satisfy

$$\hat{b} = \sum_{i=1}^{N} u_i / N \qquad (4.1)$$

and

$$\chi(\hat{k}) - \log(\hat{k}) = \frac{1}{N} \sum_{i=1}^{N} \log u_i - \log\left(\sum_{i=1}^{N} u_i / N\right) \qquad (4.2)$$

where

$$\chi(x) = \frac{d}{dx} \log \Gamma(x)$$

The function on the left-hand side of equation (4.2) was tabulated by a computer routine, and then an estimate $\hat{k}$ was formed by quadratic interpolation. For the chest X-Ray, $\hat{k}$ was evaluated as 2.82.

TABLE 4.8 OBSERVED AND THEORETICAL DISTRIBUTIONS OF TIME TO PERFORM CHEST X-RAY EXAMINATION

Time interval (minutes)	Number of patients in sample	Expected number with gamma distribution
$0 \leq t < \frac{1}{2}$	11	5.3
$\frac{1}{2}$ -	25	19.8
1 -	27	29.4
$1\frac{1}{2}$ -	36	33.5
2 -	24	29.8
$2\frac{1}{2}$ -	18	26.6
3 -	16	21.5
$3\frac{1}{2}$ -	18	16.4
4 -	13	10.2
$4\frac{1}{2}$ -	11	8.2
$t \geq 5$	24	21.8
Totals	223	222.5

The expected numbers in the sample in each of the time intervals, assuming a gamma distribution of parameters b and k, were calculated (using Pearson's Tables of the Incomplete Γ-Function and a bivariate mid-point central difference interpolation). The results are also shown in Table 4.8. A goodness-of-fit test yielded a χ^2 statistic of 15.2, distributed with 10 degrees of freedom; this is not quite significant at the 90% level. It may be seen that the sample contains rather more observations at the two ends of the time range than are expected with a gamma distribution. However the adoption of this distribution in the later parts of this work does not appear wholly unreasonable. This confirms the findings of Bailey and Fraser, amongst others, who also used this distribution in theoretical and simulation studies of hospital clinics.

4.6.3 Variation of Service Time with Age and Other Factors

The above calculations involve only the data concerned with the service time of each patient. When patients are classified by age, origin and mobility, other patterns emerge. Figure 4.4 gives a plot of the estimated mean and standard deviation of service times of patients grouped by age into ten-year intervals, and Table 4.9 gives the estimated means and variances, and the numbers in each age group.

Old people and young children sometimes presented additional difficulties when examined. Often they did not understand, or took longer to comply with the directions of the staff, particularly with respect to the positioning of the body relative to the plate, expanding the lungs, or keeping stationary during the exposure. Frequently in such cases a plate would be wasted by a patient movement just at the vital moment of exposure. Normally the radiographers would "hide" from radiation scatter behind a protective shield near

FIGURE 4.4 ESTIMATED MEANS AND STANDARD DEVIATIONS OF TIMES TO PERFORM A CHEST X-RAY: PATIENTS GROUPED INTO 10-YEAR INTERVALS OF AGE

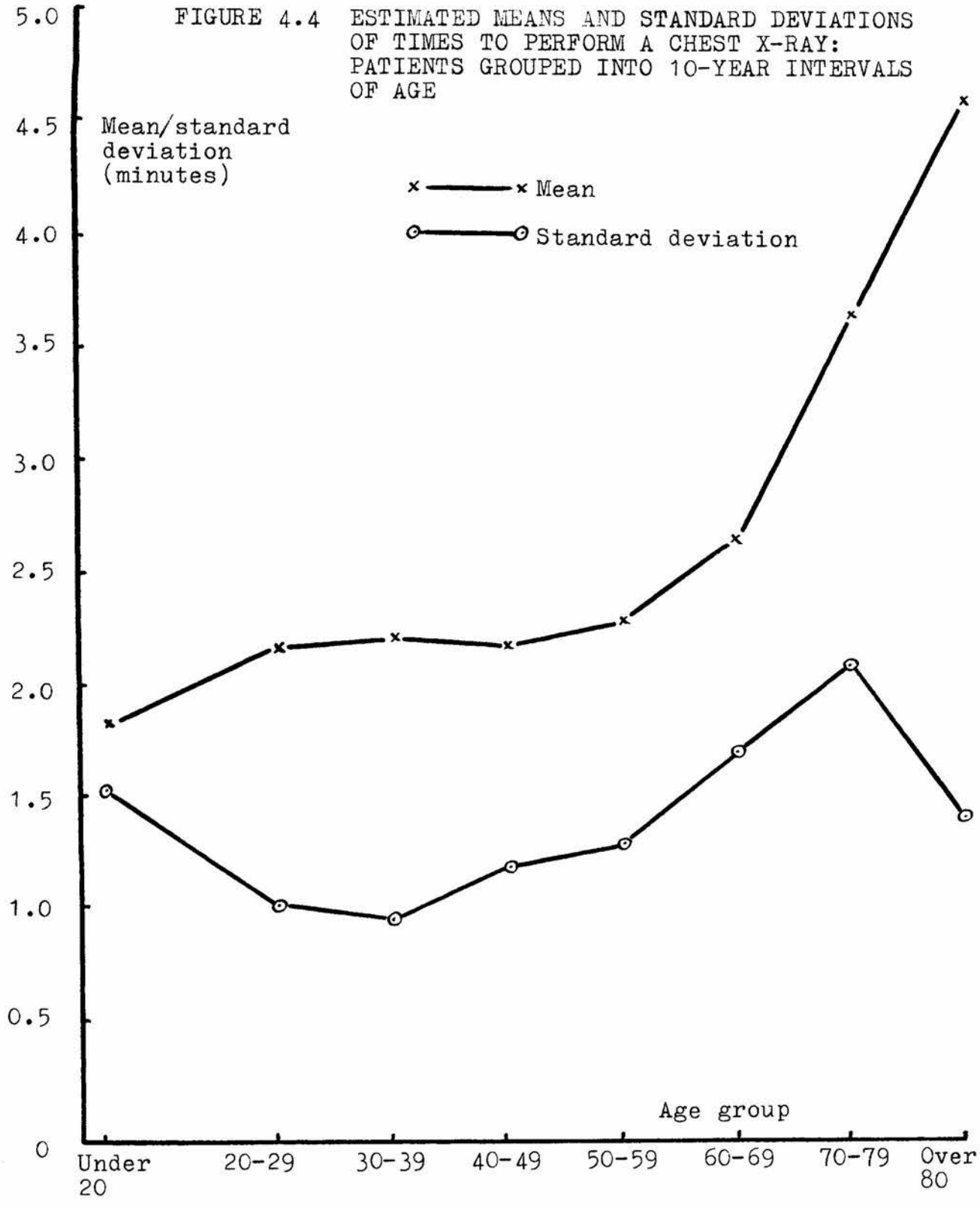


TABLE 4.9 ESTIMATED MEAN AND VARIANCE OF TIMES FOR CHEST X-RAY OF PATIENTS GROUPED BY AGE

Age Group	Number in sample	Estimated mean (minutes)	Estimated variance (minutes^2)
0 - 9	1	3.75	-
10 - 19	21	1.73	2.23
20 - 29	10	2.15	1.00
30 - 39	25	2.21	0.89
40 - 49	41	2.19	1.39
50 - 59	38	2.27	1.63
60 - 69	60	2.64	2.82
70 - 79	48	3.59	4.30
80 - 89	11	4.52	1.93
90 -	1	4.83	-
Total	256	2.56	3.04

the control panel during the exposure; in difficult cases, however, it was possible for them to encourage or hold the patient in position during the exposure by donning a protective rubber garment. As this gave only partial protection from radiation, most workers did not like to be repeatedly exposed in this manner. It was sometimes possible for a relative to hold the patient under staff direction, but cases of this type almost invariably took much longer than normal to complete.

These considerations are reflected in Figure 4.4 and Table 4.9. We may note a substantial increase in the mean service time for the over-60 age groups. There is also a tendency for the variance to increase with age, other than for very old or very young patients.

Table 4.10 shows the estimated service time means and variances for various categories of patient examined on the conventional machine, and the corresponding information for the Odelca machine. We observe that the service time mean and variance both increase substantially with decreasing mobility and fitness of the patient; for example trolley patients take almost twice as long on average as walking patients, and are also far more variable. Also evident are the lower mean times on the Odelca, and the very low associated variances; this specialised machine could be regarded as one example of a "labour-saving device" as recommended by the Nuffield team.

Table 4.11 shows the total number of patients in the sample, classified into age groups, by mobility and origin. We note that in the age distribution for inpatients a higher proportion of cases in the older age groups is shown than for outpatients; also the age distribution of patients examined on the Odelca is more concentrated over the younger age groups than in corresponding

TABLE 4.10 ESTIMATED MEAN AND VARIANCE OF TIMES FOR CHEST X-RAY OF VARIOUS PATIENT GROUPS

(a) Conventional machine

Patient Group	Number in sample	Estimated mean (minutes)	Estimated variance (minutes^2)
Inpatients:			
Walking	32	2.61	2.41
Wheelchair	68	3.38	3.86
Trolley	18	4.67	4.71
Total	118	3.37	3.96
Outpatients:			
Walking	101	1.89	1.20
Wheelchair	2	3.54	-
Trolley	2	5.17	-
Total	105	1.98	1.83
In- and outpatients	223	2.72	3.23

(b) "Odelca" machine (all patients walking)

Inpatients	13	1.88	0.15
Outpatients	20	1.21	0.34
Total	33	1.46	0.27
Both machines	256	2.56	3.04

TABLE 4.11 TOTAL NUMBER OF PATIENTS IN SAMPLE IN VARIOUS CATEGORIES

	Age group								
	0-19	20-29	30-39	40-49	50-59	60-69	70-79	80-	Total
(a) Conventional machine									
Inpatients:									
Walking	3	1	2	4	6	8	9	0	33
Wheelchair	2	1	1	6	11	14	22	10	67
Trolley	0	0	0	3	0	3	10	2	18
Total	5	2	3	13	17	25	41	12	118
Outpatients:									
Walking	6	5	12	23	19	28	7	0	100
Wheelchair	0	0	0	0	0	3	0	0	3
Trolley	0	0	0	0	1	1	0	0	2
Total	6	5	10	25	20	32	7	0	105
In- and outpatients	11	7	15	36	37	57	48	12	223
(b) "Odelca" machine									
Inpatients	1	0	6	3	1	2	0	0	13
Outpatients	10	3	4	2	0	1	0	0	20
Total	11	3	10	5	1	3	0	0	33
Both machines	22	10	25	41	38	60	48	12	256

categories for the conventional machine. The precise relationship of service-time with the factors age, origin and mobility might well form the basis of a further study. The interactions between all the factors are important; it might also be noted that very roughly the variances shown in Table 4.10 are approximately proportional to the means - this could perhaps yield a useful transformation of the data. In fact simple linear regressions of service times on age were calculated with the untransformed data in each of the mobility categories for inpatients, and also walking outpatients. The largest slope coefficient turned out to be only 25% larger than the smallest, and so an assumption of parallelism might be appropriate. One would of course have to use indicator variables for the qualitative regressors mobility and origin.

The last factor to be discussed having a possible influence on service times is the sex of the patient. Table 4.12 shows the age distribution of patients in the sample, classified also by mobility, and also the respective estimates of mean and variance of the service times. There seems to be no clear evidence of any difference in the variances for each sex, and numerical differences in the estimates are small; therefore simple t-tests were made of the hypotheses that the service time means in each patient group were the same for both sexes. Tests were also made for the equality of variance in comparable groups using an F-statistic. Table 4.13 gives the results of these tests, showing the values of the test statistics, the degrees of freedom, and the indicated significances at the 90% and 95% levels. The results seem to indicate no substantial difference between the two sexes, and the same distribution will be assumed for both in the rest of this work.

TABLE 4.12 TOTAL NUMBER OF PATIENTS IN SAMPLE GROUPED BY SEX AND MOBILITY WITH ESTIMATED MEANS ($\bar{x}$) AND VARIANCES (s^2) FOR CHEST X-RAY TIME

	Age Group										
	0-19	20-29	30-39	40-49	50-59	60-69	70-79	80-	Total	$\bar{x}$	s^2
Male:											
Walking	8	4	8	11	14	24	6	0	75	2.16	1.85
Wheelchair	1	0	1	3	6	10	8	2	31	3.20	1.73
Trolley	0	0	0	2	1	2	7	2	14	4.17	1.56
Total	9	4	9	16	21	36	21	4	120	2.66	2.26
Female:											
Walking	11	5	16	19	11	18	12	0	92	1.88	1.26
Wheelchair	2	1	0	6	5	6	13	5	38	3.53	2.42
Trolley	0	0	0	0	1	0	2	3	6	6.01	3.29
Total	13	6	16	25	17	24	27	8	136	2.52	2.74

TABLE 4.13 SIGNIFICANCE TESTS INVOLVING PATIENT GROUPS OF OPPOSITE SEX

(a) Tests for equality of service time means

Patient Group	Value of t statistic	Degrees of freedom	Significance at levels: 90%	95%
Walking	1.10	165	No	No
Wheelchair	0.65	67	No	No
Trolley	1.84	18	Yes	No
All patients	0.71	254	No	No

(b) Tests for equality of service time variances

Patient Group	Value of F statistic	Degrees of freedom	Significance at levels: 90%	95%
Walking	1.47	74, 91	Yes	No
Wheelchair	1.40	37, 30	No	No
Trolley	2.11	5, 13	No	No
All patients	1.21	119,135	No	No

4.6.4 Administrative Policy of Patient Segregation

Even if it were possible to estimate the main effects of the factors age, mobility and origin on the service time, it is important to consider the possible policy changes to be made as a result. Essentially the only control the X-Ray department has on the work load is by giving some patients appointments at appropriate times. Also in the long term it may be possible to encourage clinics and other sources to send their patients at particular times of day. Thus policy decisions of this kind by the administration consist solely in the sequencing of work as far as possible, to give optimum working conditions and efficiency.

By these considerations, we are driven to investigate the effects of sequencing some of the patients according to one or more of the factors acting on service times. In particular we must look at the behaviour of the system when it works either in sessions of groups of segregated patients, each having the same factor characteristics, or in sessions made up of a few subsessions of homogeneous groups as before.

It is difficult to envisage an operating policy which segregated patients according to the mobility factor alone; often it is not known how fit a patient is before his arrival, but such a regime would clearly be unreasonable and unworkable anyway. When it is known in advance that, for example, a trolley patient is going to prove much more difficult than usual, a form of segregation might be employed by examining that patient towards the end of a clinic. Patients who have a service time much longer than average cause much less congestion if treated at this time, rather than at the start of a session, causing all subsequent patients to have greatly increased queueing times.

Similar considerations apply to segregation by age. Again this would be

rather impractical to operate totally, but exceptionally old patients might profitably be timetabled for the end of work periods.

Finally we are led to consider segregating by the origins of patients. In practice, of course, it would not be possible to have a timetable of sessions consisting entirely of patients from a small number of specified sources, but the general idea is worth pursuing. Table 4.10 showed substantially lower mean service times for out-patients, and even a simple grouping of patients into two sets in this manner might yield results. Of course the origin of the patient is not a "reason" for the higher service time, as are the patient's age and mobility. Table 4.11 indicated the different age-mobility structures for in- and out-patients for chest X-Rays; we extend the idea by considering other examinations.

Figures 4.5, 4.6, 4.7 and 4.8 show the distributions of age of patients of various origins for all examinations; this information was extracted from the random sample of the hospital records. The inpatient distribution of Figure 4.5 is dominated by a majority of patients in the 50-79 age bracket; in fact 65% of the inpatients were over 60, which were the ages of substantially higher mean service times for the chest X-Ray. There is also a much smaller mode in the 30-39 age group. Both the outpatient distributions for casualty and other sources demonstrate bimodality; the distribution of Figure 4.6 of patients mostly from clinics has peaks in the 20-29 and 60-69 age groups; in the casualty distribution, the lower mode is somewhat more prominent, with the age bracket 10-29 containing 52% of the input from this source. Both outpatient distributions show a local minimum in the group 30-39.

FIGURE 4.5 AGE DISTRIBUTION OF INPATIENTS

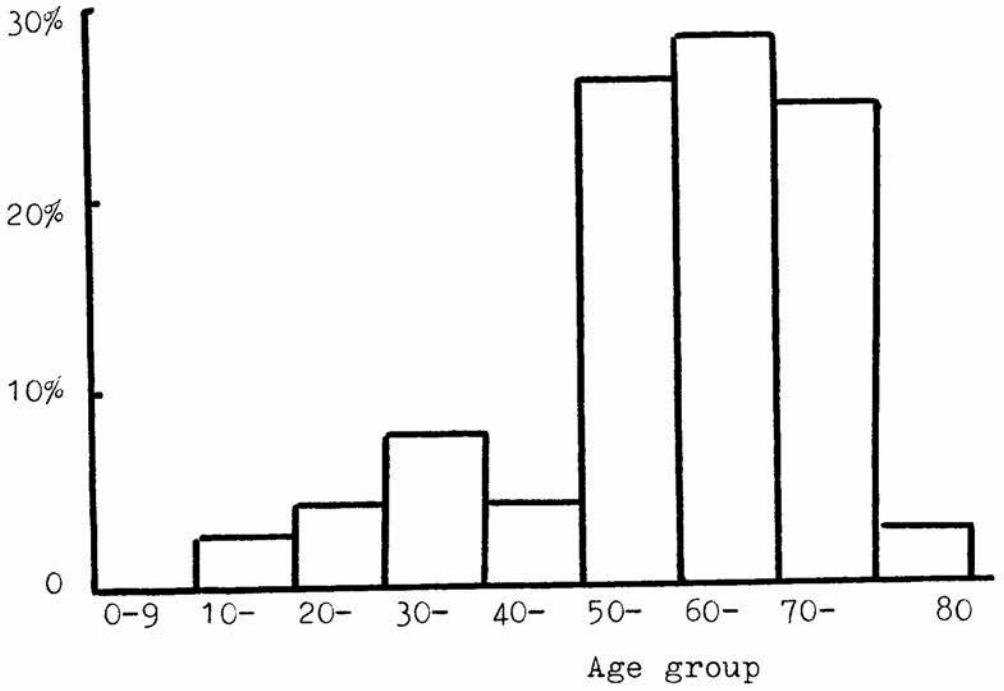


FIGURE 4.6 AGE DISTRIBUTION OF OUTPATIENTS (EXCEPT CASUALTY)

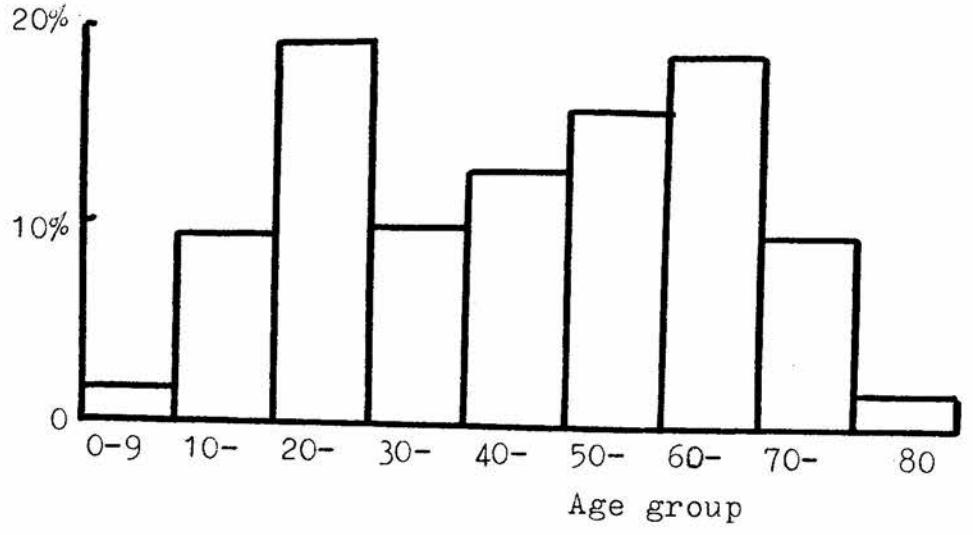


FIGURE 4.7 AGE DISTRIBUTION OF ACCIDENT AND EMERGENCY PATIENTS

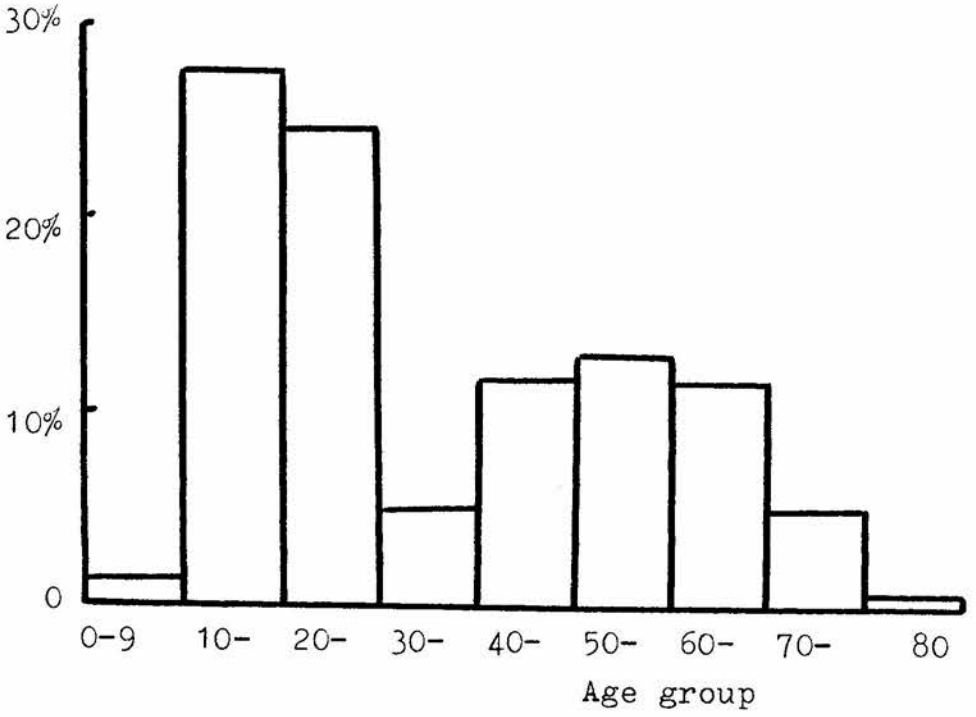
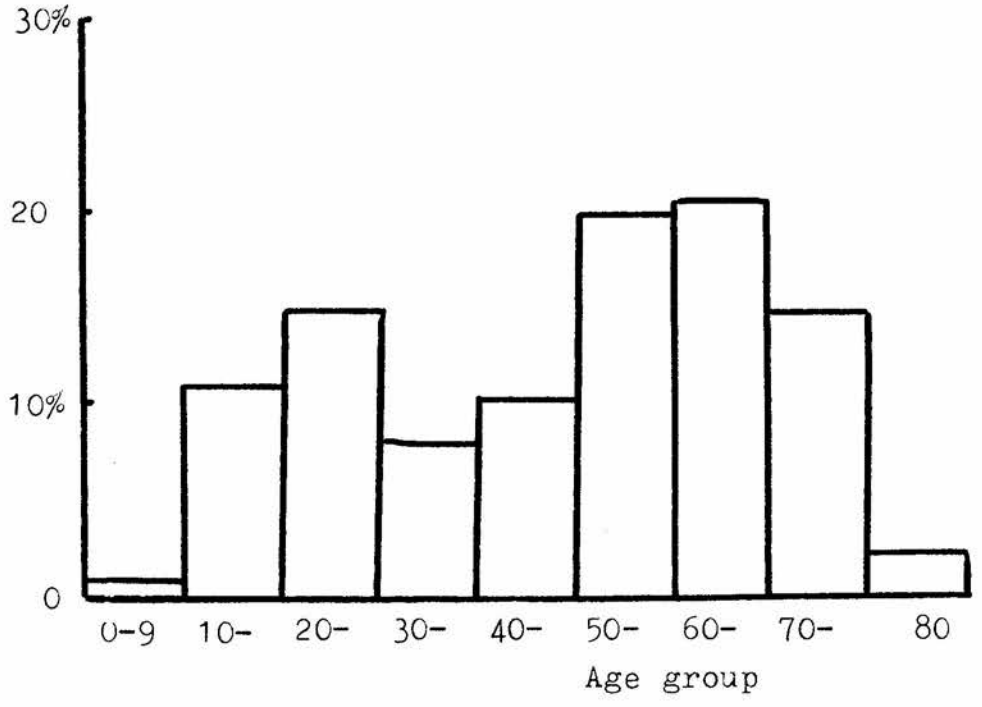


FIGURE 4.8 AGE DISTRIBUTION OF ALL PATIENTS



Thus apart from isolated patients, "origin" is the only factor by which we can usefully segregate patients in a practicable way, and it is clear that even a relatively simple grouping of patients into in- and outpatient sets would yield a considerably higher degree of homogeneity in the input stream. In any particular hospital, it should be possible to extend this idea by segregating, say, patients from geriatric wards or intensive care units, or any class of patient known to have a longer or more variable service time.

The following chapter, which develops a theoretical model for a clinic, and the subsequent chapter describing further studies using simulation techniques, use this idea of segregation, and attempts are made to quantify some effects of policy changes of the type described earlier.

4.7 Accident and Emergency Department (Casualty)

The casualty department was one area of the hospital which did not fit easily into the general pattern of work observed elsewhere. In view of the special circumstances there, it was decided to make a small survey of its work.

By the very nature of this department, X-Ray work might arrive at any hour of the day or night, and no cases were "predictable". A sample was made of the total work of the department for a two week period in July 1971 from the department day-book. From this source it was possible to obtain the numbers of patients arriving in the time intervals 8.30 a.m. to 5 p.m., 5 p.m. to 9 p.m., and 9 p.m. to 8.30 a.m; an exception was Sunday when the staff shift ended at 4.30 p.m. instead of 5 p.m. Table 4.14 shows the values obtained in this sample.

TABLE 4.14 NUMBER OF PATIENTS EXAMINED IN CASUALTY DEPARTMENT DURING A TWO WEEK PERIOD, FOR VARIOUS HOURS OF THE DAY

Period	Day							
	Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Total
8.30 a.m. to 5 p.m.	94	132	117	105	99	113	87	747
5 p.m. to 9 p.m.	34	33	51	47	42	49	38	294
9 p.m. to 8.30 a.m.	36	45	42	44	41	61	68	337
Total	164	210	210	196	182	223	193	1378

A test was made of the hypothesis that the total demand was constant over the days Monday to Saturday; the figures indicated no significant difference at the 90% level from uniformity, with a value of 5.36 for the statistic distributed as χ^2 with 5 degrees of freedom. The combined figures for the 8.30 a.m. to 9 p.m. period also showed no significant difference from uniformity at the 90% level, with $\chi^2_5 = 8.40$. However the night period totals were significantly different at the 95% level, with $\chi^2_5 = 12.97$; this probably reflects the increased demand on Friday and Saturday nights. Staffing arrangements were the same on these nights as during the week, despite the fact that this demand pattern had long been recognised, if only qualitatively.

The casualty staff was partly responsible for carrying out the so-called "portable" or "mobile" examinations with a small machine for patients too ill to be moved from their wards. Some of these were routine, performed every night by casualty staff who were normally responsible for all such examinations at night. Because of local staffing arrangements, casualty staff also did all the mobile examinations during the day on Sunday. Mobiles were frequently performed in intensive care wards as routine or emergency, and sometimes pre-operatively, particularly on Sundays for the last category. An emergency operating theatre also gave a small amount of work (see Table 4.1), again usually at night. Mobiles were the only class of examinations which showed any marked difference from the overall distribution of work throughout the week, the rest showing much the same pattern as in the main department.

The service time distribution followed in general the same pattern as for the main department, with approximately the same age weightings. However there did appear to be a rather higher proportion of outliers, or large

observations, as compared to the main department; these had the effect of increasing the mean service time by 10% to 20% over those observed on similar machinery in the main department. These long service times often occurred when the patient was in pain and difficult to position for the X-Ray, and sometimes when the patient was in shock after an accident, becoming hysterical or unresponsive to instructions from the staff. The same difficulties were evident in keeping the patient stationary as in the chest room work in the main department described earlier, only in casualty it was less often possible to use relatives to hold the patient and aid the staff, as they themselves were often in a shocked condition after an accident.

As may be seen from the random sample of the departmental archives (Tables 4.3, 4.4 and 4.5), the casualty department handled a high proportion of the extremity work, and rather less chest work than elsewhere. For the resources available, the department worked at a rather slower rate on average than the other main department rooms, and considerably more slowly at night. However average rates are not too meaningful in this situation, as patients would commonly arrive in batches, often requiring fairly technical examinations, and the staff of this department, by necessity in constant readiness, appeared to be working at least as hard as their counterparts in other departments.

4.8 Summary of Problem Areas in the Royal Infirmary

This section outlines in general terms the particular areas of the Royal Infirmary which might be investigated with a view to efficiency improvements.

At this hospital, only the chest room showed an average patient occupancy of over 40%, but it must be remembered that the Royal Infirmary is a large teaching hospital handling cases which would be rarely dealt with at

an equivalent ordinary hospital. Several rooms were used only by senior radiologists for specialist examinations, and even the main bulk of work contains a proportion of technical work unlikely to be seen often at other hospitals.

On consideration of these low occupancies of the facilities, and the substantial periods of the day when little or no work is done, it might appear that the department was working well within its maximum capacity. However this situation does seem to have been bought partly at the expense of the patients who frequently have to queue and wait excessively at peak periods.

The single step which might perhaps cause the greatest improvement in efficiency would seem to be a more even distribution of the work load throughout the working day. Other clinics might be encouraged to send X-Ray cases early in the session, particularly when it is known in advance that a particular patient will definitely require a visit to the X-Ray department; also, attempts could be made to deal with ward-patients at off-peak times whenever possible. Many conflicting objectives of policy within the hospital framework exist here, but in principle other units of the hospital would surely co-operate to achieve a better service for their patients.

Patients who seemed particularly unfortunate at the Royal were those needing multiple examinations. It was the exception for any special allowance or priority to be made in queueing for the second examination, and by queueing two or three times they would often spend several hours in the department for reasonably simple examinations. As an improvement, it might be possible to give higher priority to such patients at the second and subsequent queueing stages. This is perhaps not quite so important for inpatients who are

frequently not in a hurry to leave for other "engagements"; indeed they often enjoy a visit to another department as a change from the ward routine. As noted in section 4.23, time savings might be made by performing combination examinations in a flexible order.

This last is but one example of having a flexible approach to and co-ordination of the activities of different areas of the hospital. A central, trained receptionist or nurse able to direct the work to free areas of the department would clearly be an advantage, and improve conditions for both staff and patient.

Other difficulties were noticed in the portering schedules. Firstly the porters were unable to get the work flow going smoothly for some time at the beginning of the day because of other peripheral tasks, such as delivering completed diagnoses. Also there seemed to be communications problems between the wards and the X-Ray departments - a porter would sometimes go to collect a patient for whom an X-Ray had been requested, only to find he had been discharged from the hospital. Bottlenecks developed from time to time at the processing and diagnostic stages, but these seemed to stem usually from an overload at one of the peak patient arrival periods. Serious delays and congestion were also observed when an old or seriously ill patient took a very long time to examine, or when patients did not receive clear directions from the staff and were consequently unprepared.

It seemed that the patient queueing and doctor idle-time issue was at the root of many of the difficulties outlined, and it was decided to make this one of the main aspects of the study. This is developed in the following chapters using the solution of a general class of queueing models, and simulation results.

5. Mathematical Models of a Clinic Queueing System

5.1 Introduction

The literature concerning stochastic processes, and queueing theory in particular, has grown immensely over the last twenty years, despite the fact that these subjects are relatively new even compared to statistics as a whole. A large proportion of the recent research in this area has been in investigating complex aspects of very simple models, or in deriving new and more aesthetic results for long-established classical models. This is partly because there seem to be all too few queueing systems which yield easy mathematical solutions; indeed the mathematics becomes complex and often intractable when all but the simplest systems are described, which require many assumptions and involve little or no generality.

There is a sizable gap between the mathematical queueing theory available and the analytical solution of models which operational research workers in most areas would regard as approaching an adequate description of their real situation. It is not normally possible to construct high-fidelity models of a real system which are mathematically tractable. Lee (1968) therefore puts forward the view, perhaps extreme, that progress in operational research has been slow in this area because of the reluctance of practical workers to discuss problems at a low mathematical level for fear of criticism from academics. He continues:-

> "Congestion problems still constitute a major area in which much practical operational research is done, but to discover this it is necessary to listen to the informal, unguarded conversations that take place in the relaxed hours following conference dinners."

Despite this gap between theory and practice, which will always exist to some degree, there remain many good reasons for using mathematical models.

An approximate model may give deeper insight into the general principles governing much more complex real situations, when the assumptions and limitations of the simple model are borne in mind. Also a small amount of analysis on the simpler model may yield general algebraic results which might be difficult or impossible to obtain by simulation or direct experimentation. These two methods might require extensive study over a wide range of parameters before even an approximation could be made to the underlying unknown general formula or relationship being sought. Thus a rough model is often better than a mathematically intractable one, and is certainly better than none at all.

On the other hand, when interpreting and applying the results from a model, we must not only pay due regard to the assumptions it embodies, but also be aware of the deficiencies of queueing theory as a whole in this context. Many of the assumptions made even in commonly used models may be inappropriate, particularly in human situations. There is still a lack of extensive data from queueing situations occuring in the real world in all but a few particular cases, notably telecommunications, one of the first practical applications of the subject. All too little is known of the psychology and behaviour of waiting people. For example, will they be daunted on seeing a long queue on arrival, and turn away? Under what circumstances will they move to another queue? Will they leave after waiting a certain time, without receiving service? Solvable models which take account of these considerations do exist, normally with very simple assumptions made about the customer behaviour, but as yet quantitative analyses of the empirical behaviour of people in queues are almost unknown.

If it is decided finally that a simple solvable model is too crude an approximation to the real world for the purposes of extracting the statistics of interest, other methods using a more empirical or qualitative approach are available. Simulation is one of these, and the particular advantages and limitations of this method in describing an X-Ray department will be discussed in the following chapter.

5.2 Queueing Theory Terminology

In this section some of the notations and definitions of queueing theory necessary in this study are presented.

A stream of customers (in this example, patients) arrive at a service facility (the X-Ray department), wait, and are given a service by a group of servers (radiographers or radiologists). Let customer i arrive at time t_i, and let $t_o = 0$; let $u_i = t_i - t_{i-1}$. Then $\{u_i ; i = 1, 2, \ldots\}$ is a set of observations from the same inter-arrival time distribution, dG(t). The observations may not be statistically independent; often they are and this facilitates analysis. The input-process is the complete description of the general arrival times t_i, and thus includes a specification of the function dG(t) and the rules for selecting observations from it. Customers queue in front of a number of service counters, and the rules governing the selection of customers for service are known collectively as the queue discipline.

We may wish to include in the model further ramifications to describe the behaviour of customers in the waiting stage and a number of these are described here. Balking occurs when customers arrive but decide not to wait if the queue is too large, or if a queue exists at all; reneging is the

phenomenon of customers losing patience after a time and leaving without service; customers under certain conditions may decide to move to another equivalent, but shorter, parallel queue, and this is called jockeying. Customer psychology may also influence the input process: for example, customers may learn from experience the likely general behaviour of the queue throughout the day, and may then modify their arrival time to coincide with an expected slack period.

Different classes of customer may be assigned priorities, possibly time-dependent, which affect their order of selection for service. Some priorities may be pre-emptive, implying that the customer currently receiving service may be left before service completion by the server, who then deals immediately with the new customer on arrival. (An example in the hospital is a cardiac arrest.)

Two or more equivalent parallel queues may be pooled, and the servers then select customers from the same queue. For example two doctors may deal jointly with a single queue of patients of the same type. This arrangement usually leads to shorter waits by the patients than in the separated queues.

The set of times to complete service on different customers forms the service-time distribution. This distribution is usually taken to be the same for all customers, and is here denoted by dH(u). Successive service times may not be independent, and the underlying distribution may be affected by other factors such as the queue length at any given instant, or the total number of customers served since an idle period.

Kendall's notation dG(t)/dH(u)/s, which denotes the queueing system with inter-arrival time distribution dG(t), service-time distribution dH(u),

and s servers in the facility, is now widely used. We will also use some particular notations. If observations are assumed independent from the same distributions dG(t) and dH(u), then the queue is GI/G/s, where GI is general independent and G is general. Several commonly used inter-arrival distributions have special notations. We shall use the following:

(i) M. Customers arrive in a Poisson process, or at random. The inter-arrival times are independently exponentially distributed and

$$dG(t) = \frac{1}{a} e^{-t/a}\, dt \qquad (0 < t < \infty)$$

(ii) D. Customers arrive at regular intervals a, and

$$G(t) = \begin{cases} 0 & t < a \\ 1 & t \geq a \end{cases}$$

(iii) E_k The intervals between arrivals are made up of the sum of k independent exponential variables with the same mean b/k. The resulting distribution is the Erlangian with k phases and mean b and

$$dG(t) = \left(\frac{k}{b}\right)^k \frac{e^{-kt/b}\, t^{k-1}}{(k-1)!}\, dt$$

k = 1 corresponds to the exponential distribution, or (i). If $k \to \infty$ in such a way that b/k = constant, then the limit distribution is regular, or (ii). If the distribution dG(t) has the same form as the Erlangian, but k is not integral, then it is one of the Pearson Type-III curves. The case of integral k is much simpler to deal with analytically

because we can always decompose the distribution into independent exponential phases. These values also characterise the behaviour of the more general Pearson III curves, and therefore the approach adopted here has been to use only integer values of k.

M and E_k will also be used in this work as service-time distributions.

We say a queueing system is in equilibrium or a "steady-state" if its state probabilities and state transition probabilities are time independent. This can be when the system has been running for a very long time and has settled down to a behaviour independent of any initial parameter values, and none of the governing factors is changing. We distinguish between waiting time, which is the time elapsing between a customer's arrival and completion of service, and queueing time, which includes only the time spent by customers in a line prior to starting service.

We now turn to the problem of building an appropriate mathematical model of the real situation, an X-Ray department.

5.2.1 The Input System

In this section we attempt to describe the pattern of arrival times occurring in practice at an X-Ray department by a theoretical distribution. We have seen in the previous chapters that there are two main classes of patient when considering this aspect of the analysis, those arriving with appointments and those without.

Several patients may be given the same appointment time, a practice known as block booking. If we take the size of a batch as m, and assume the appointment intervals b are constant, then when m = 1 the subgroup of appointment arrivals forms a stream with regular input process, or D in Kendall's notation. For general m, and constant b, we will denote by D_m the stream of arrivals in batches of m at regular time intervals b.

For appointment patients, the distribution of {time of arrival - time of appointment} is known as the lateness or unpunctuality distribution. When patients are unpunctual the behaviour of the queueing system may be affected, and much of Pike's work (1963a) was concerned with the investigation of this aspect of appointment schemes; the relevant parts of this work are also given by White and Pike (1964). Simulation results of a queueing system with unpunctual patients were compared with theoretical results of a system with punctual patients. In general terms it was found that the two sets of results were only slightly different. White and Pike also considered the problem of "missing" patients who fail to arrive at all; by decreasing the appointment interval to give the same expected traffic intensity, it was found that the queueing system had very similar characteristics to that with no missing patients. In the rest of this work we shall assume that all appointment patients arrive, and are punctual. However the doctor's punctuality at the beginning of the clinic session has a strong influence on the subsequent behaviour of the system, and this variable is included in the main theoretical model of this chapter beginning in section 5.4, and in the simulations of chapter 6.

Patients arriving without appointments originate from many sources. Often a patient attending a specialist clinic, such as the Ear Nose and Throat clinic for example, will be sent for X-Ray, and his treatment will be continued later on the basis of the information derived; thus these patients normally arrive in the X-Ray department without prior warning. Casualty and emergency patients do the same, and need immediate attention; non-appointment patients may also originate from the wards of the hospital. Each of these sources may be regarded as having some sort of underlying renewal process, and when the parameters of each remain constant and their outputs of patients are pooled into a common arrival stream at X-Ray, then we may regard the whole complex as having many superimposed renewal processes. In the limit as the number of sources becomes large, the pooled output is governed by a Poisson process (Cox and Smith 1954); in other words the arrival stream may be regarded as having a random form with exponentially distributed inter-arrival times.

Obvious difficulties arise when using this approach in the hospital situation, as the parameters of each arrival stream are far from constant throughout the day. For example, suppose an out-patient clinic sending some patients for X-Ray has a morning session from 9 a.m. till 12.30 p.m. Most of these out-patients require at least some attention, if only administrative, at their specialist clinic before being sent for X-Ray. Also towards the end of the session it becomes impractical to send any more patients and leave enough time for them to return to the clinic for further treatment in the same session. Thus we might expect a variable rate of arrivals from such a clinic, which starts at

a low level at 9 a.m., climbs to a peak and decreases once more towards 12.30 p.m. In fact this does occur in many clinics, and often hospital X-Ray departments have a well-defined "peak" of this type of arrival at 10.30 or 11 a.m., and another at 2.30 or 3 p.m. for the same reasons in the afternoon sessions. The composition of arrival streams may also vary, because of some priority system within clinics. The pattern of emergency arrivals is also not constant throughout the day (cf. section 4.7 and "Towards a Clearer View", Ch. 5). Lastly there may be consistent differences in the parameters between days of the week or parts of the year, and the whole pattern may have some underlying trend.

Despite all these comments, we will assume for the present that the overall rate of arrivals of non-appointment patients is constant for the duration of any one session of the X-Ray department, that the composition of this stream is constant, and that the input process is random, or M.

In a real hospital, we will have a mixture of appointment and non-appointment arrivals in some ratio which is possibly time dependent. We will denote a mixture of random arrivals, and regular arrivals in batches of m by $(M + D_m)$ in the input process.* It should be noted that with such an input, successive inter-arrival times are not statistically independent. We denote the ratio of regular arrivals

---

*It should be pointed out that the conventions D_m and $(M + D_m)$ are my own, and not in widespread use. They do, however, adhere to the Kendall system.

by appointment to the randoms without appointments, by r, which we will assume to be constant for the duration of any one session of the X-Ray department. As $r \to o$, the input-process becomes M, and as $r \to \infty$, it becomes D_m.

5.2.2 Queue Discipline and Service Mechanism

We will assume a FIFO (first in, first out) queue discipline; in other words patients are served in order of arrival, irrespective of type. This is clearly an inadequate description of a real system which has some emergency patients with high or pre-emptive priorities, but this situation will be examined later. All members of a regular batch arrival have equal priority, and are served successively in random order.

For the service-time distribution we will assume a Pearson-III or gamma type; we will consider only integer values of k for simplicity of analysis. The distribution is equivalent to a k-phase Erlangian, or a scale-modified χ^2 with 2k degrees of freedom. When $k = 1$ the distribution is exponential. The Erlangian has been shown to adequately describe the distributions actually observed in a wide range of situations in clinics, and also gave a satisfactory fit to data of Chapter 4.

When no patients arrive by appointment, we have a system yielding model (I): M/M/1 or (II): $M/E_k/1$ depending on the service distribution used. Model (III): $E_{k_1}/E_{k_2}/1$ is useful as particular values of k_1 and k_2 yield M and D for the input process. Models (IV): D/M/1 and (V): $D/E_k/1$ are used when there are no random arrivals; to generalise to block bookings we use (VI): $D_m/E_k/1$. Finally the most general models used here, involving a mixed input stream, are (VII): $(M + D_m)/M/1$ and (VIII): $(M + D_m)/E_k/1$.

5.3 Solutions of Models

5.3.1 Model (I): M/M/1

Model (I) is the best known queueing model and is in fact called the simple queue. The equilibrium queueing time distribution and the distribution of the number in the system are reasonably easy to derive, and many time-dependent properties are also known. As this is a particular case of some of our later models, we will postpone the presentation of the results until comparisons with this system are required.

5.3.2 Model (III): $E_{k_1}/E_{k_2}/1$

A general approach to a wide class of queueing problems, of which this constitutes one, was derived by Smith (1953), and this particular application is developed here. Alternative procedures for model (II): $M/E_k/1$ and also $E_k/M/1$ may be derived by writing the steady-state equations for each.

Let G(t) be the inter-arrival distribution, and H(u) the service time distribution, K(x) the distribution of $x = u-t$, and F(v) the queueing time distribution of v. Under fairly general independence and equilibrium conditions, Lindley (1951) obtained a Weiner-Hopf integral equation for the queueing time distribution:-

$$F(v) = \int_0^\infty K(v - u)dF(u)$$

Following Cox and Smith (1967) in a particular case of the original work, we assume here that the inter-arrival intervals can be regarded as the sum of p independent exponential variables, with parameters $a_1, a_2, \ldots . a_p$. In particular p = 1 gives random arrivals. We also have to assume G(t) continuous, with Laplace transform

$$G^*(s) = \int_0^\infty e^{-st} \, dG(t)$$

$$= \frac{a_1 \cdot a_2 \cdot \ldots \cdot a_p}{(a_1 + s)(a_2 + s) \ldots (a_p + s)} \qquad (Re(s)>0)$$

If we apply the operator $(1 - 1/a_1)(1 - 1/a_2)\ldots\ldots(1 - 1/a_p)$ to Lindley's equation and take Laplace transforms, we obtain

$$P + F^*(s) = \frac{Q_p(s)}{\{G^*(-s)\}^{-1} - H^*(s)}$$

where * denotes a transform and $Q_p(s)$ is an unknown polynomial of degree p in s. P is the probability of zero queueing time.

We write

$$G^*(s) = \{g_p(s)\}^{-1}$$

where $g_p(s)$ is a polynomial in s of degree p. Also for the cases we are considering, we may write

$H^*(s) = \{h_k(s)\}^{-1}$, where $h_k(s)$ is a polynomial of degree k in s. Thus

$$P + F^*(s) = \frac{Q_p(s).h_k(s)}{h_k(s).g_p(-s) - 1} \qquad (5.1)$$

The denominator of (5.1) is of degree k + p, and therefore has k + p real roots; one of these roots is at zero as the transform of any continuous distribution function at zero is unity. We also note that

$$|F^*(s)| = \left|\int_0^\infty e^{-sv} dF(v)\right|$$

$$\leq \int_0^\infty |e^{-sv} F(v)| dv$$

$$\leq \int_0^\infty |F(v)| dv \qquad \text{(for } Re(s)>0\text{)}$$

$$= \int_0^\infty F(v)dv \quad < \infty$$

For $F^*(s)$ to remain finite in equation (5.1), $Q_p(s)$ must be of the form $(s - s_1)(s - s_2)\ldots. (s - s_{p-1})s$. Therefore we must have exactly $(p - 1)$ zeros z of $G^*(- s)^{-1} - H^*(s)$ with $Re(z)>0$, and it follows that there are exactly k zeros with $Re(z)<0$. We denote these by $z_1, z_2, \ldots. z_k$. Now write $h_k(s).g_p(- s) - 1$ as

$$\delta\ (s - z_1)(s - z_2)\ldots.. (s - z_k)Q_p(s)$$

where δ is some constant. Thus

$$P + F^*(s) = \frac{h_k(s)}{\delta(s - z_1)(s - z_2)\ldots. (s - z_k)}$$

We let $s \to 0+$, and so

$$P + F^*(0) = \frac{1}{\delta \prod_{i=1}^{k} (-z_i)} = 1$$

Hence

$$P + F^*(s) = \frac{h_k(s)}{(1 - \frac{s}{z_1})(1 - \frac{s}{z_2}) \cdots (1 - \frac{s}{z_k})} \tag{5.2}$$

We now take $H(u)$ as the k-phase Erlangian distribution, and we let $k/b = \sigma$. Then

$$H^*(s) = \left(\frac{\sigma}{\sigma + s} \right)^k$$

and

$$h_k(s) = \left(\frac{\sigma + s}{\sigma} \right)^k$$

Hence

$$P + F^*(s) = \frac{\left(\frac{\sigma + s}{\sigma} \right)^k}{\prod_{i=1}^{k} \left(1 - \frac{s}{z_i} \right)} \tag{5.3}$$

We write

$$h_k(s) = \prod_{i=1}^{k} (1 - \frac{s}{\lambda_i}) ,$$

where the λ_i's are the roots λ of $h_k(s)$ with $\mathrm{Re}(\lambda) < 0$; we let $s \to \infty$ in equat-

ion (5.2) and so

$$P = \prod_{i=1}^{k} \left(\frac{z_i}{\lambda_i} \right)$$

With the Erlangian as the service time distribution, $\lambda_i = -\sigma$ for $i = 1,2,\ldots k$, and so

$$P = \frac{\prod_{i=1}^{k} (z_i)}{(-\sigma)^k} \quad ,$$

where the z_i's are the roots of

$$\left(\frac{\sigma + s}{\sigma} \right)^k \left(\frac{a_1 - s}{a_1} \right) \left(\frac{a_2 - s}{a_2} \right) \cdots \left(\frac{a_p - s}{a_p} \right) - 1 = 0 \qquad (5.4)$$

with $Re(z_i)<0$. Knowing the z_i's one can in principle invert (5.3) to derive the queueing time distribution. When the inter-arrival distribution is also Erlangian, with constant parameter a for all p phases, (5.4) simplifies to

$$\left(\frac{\sigma + s}{\sigma} \right)^k \left(\frac{a - s}{a} \right)^p - 1 = 0$$

We may extend this model by introducing the possibility of technical deficiency when an X-Ray picture is taken. Let us ignore the time for processing and reporting on a film, and let us suppose that pictures are taken repeatedly until a satisfactory plate is obtained. We assume that at each stage the service time has the same underlying k-phase Erlangian distribution. We denote the probability that j pictures are required to obtain a satisfactory

report by c_j. This regime is equivalent to saying that with probability c_j, the total service time can be regarded as having a kj-phase Erlangian distribution of mean kb. Thus

$$H^*(s) = \sum_{i=1}^{\infty} c_i \cdot \left(\frac{\sigma}{\sigma + s} \right)^{ki} ,$$

which in general is not the reciprocal of a polynomial in s, and so the previous method fails. We may, however, evaluate the first two moments of H(u) directly:-

$$E(u) = \sum_{i=1}^{\infty} c_i \cdot ib$$

and

$$E(u^2) = \sum_{i=1}^{\infty} c_i \cdot \frac{kb^2 (ki + 1)}{i}$$

Then from Pollaczek's formula we may obtain the expected waiting time ω as

$$E(\omega) = \left\{ 1 + \frac{\rho(1 + c^2)}{2(1 - \rho)} \right\} \sum_{i=1}^{\infty} c_i \cdot ib$$

where ρ is the traffic intensity and c is the coefficient of variation of service times. However the mean wait is not a very useful single statistic because of the long "tail" of the waiting time distribution.

We now make the further assumption that at each stage there is a constant probability θ that a picture will be satisfactory; then the probability of requiring exactly j pictures is

$$c_j = (1 - \theta)^{j-1} \cdot \theta .$$

We now have

$$H^*(s) = \theta\left(\frac{\sigma}{\sigma+s}\right)^k + \theta(1-\theta)\left(\frac{\sigma}{\sigma+s}\right)^{2k} + \theta(1-\theta)^2\left(\frac{\sigma}{\sigma+s}\right)^{3k} + \ldots$$

$$= \frac{\theta\left(\frac{\sigma}{\sigma+s}\right)^k}{1-(1-\theta)\left(\frac{\sigma}{\sigma+s}\right)^k}$$

$$= \frac{\theta\sigma^k}{(\sigma+s)^k - (1-\theta)\sigma^k}$$

This is the reciprocal of a polynomial in s (H(u) is thus a K_n distribution in Smith's notation), so we may apply the above method. As before we write

$$F^*(s) = \frac{\prod_{j=1}^{k}\left(1-\frac{s}{\lambda_j}\right)}{\prod_{j=1}^{k}\left(1-\frac{s}{z_j}\right)}$$

where the λ_i's are the poles of $H^*(s)$, and the z_i's are the zeros of $H^*(s)$. $G^*(-s) - 1$ with $Re(z_i)<0$. After rearrangment we find the z_i's to be the zeros of

$$(a-s)^p\{(\sigma+s)^k - \phi\,\sigma^k\} - \theta\sigma^k a ,$$

where $\phi = 1-\theta$, and p is the phase number of the input process. If $p = 1$ this expression reduces further to

$$(a-s)(\sigma+s)^k + \sigma^k(\phi s - a) .$$

The values of k occurring in practice in hospital clinics are in the range

1 to 5. For such values the z_i's could be found numerically without too much difficulty. We now write $d_i = -\lambda_i$ and $b_i = -z_i$. Then

$$F^*(s) = \prod_{i=1}^{k} \left(\frac{b_i}{d_i}\right) \left\{ \frac{1 + (f_1 - \ell_1)s^{k-1} + (f_2 - \ell_2)s^{k-3} + \ldots + (f_k - \ell_k)}{\prod_{i=1}^{k} \left(1 - \frac{s}{z_i}\right)} \right\}.$$

where f_j is the sum of all possible j-products of all distinct d_i's, and ℓ_j is the corresponding sum for the b_i's. Writing $F^*(s)$ in the form $P(1 + E)$ in this way, the Laplace inversion can now be carried out by splitting E into k partial fractions, and the final queueing time distribution will be of the form:-

$$\begin{cases} \Pr(u = 0) = P \\ F(u) = \text{k weighted exponential distributions of parameters } z_i \\ \qquad\quad (u > 0) \end{cases}$$

Alternative Method

For general k_1 and k_2, an alternative method of solution is available in principle. We may consider the joint process $\{n_1, n_2\}$, where n_1 is the number of phases completed in the arrival mechanism since the last customer arrival, and n_2 is the number of phases waiting and in service at any given instant. Under appropriate transition rules, n_1 and n_2 jointly obey a Markov process, and we may write down the steady-state equations governing it. We may then sum over suitable ranges of n_1 to obtain the steady-state probabil-

ities of the number of phases in the system.

The algebra for this method is tedious, and it is doubtful if there is any advantage over the previous method using Laplace transforms, which leads directly to the queueing time distribution. Smith's method does require the evaluation of the z_i's, which are complex, and must usually be found in practice by iteration. When we consider particular values of k_1 and k_2, this second method may yield a simpler evaluation of particular statistics of the system.

5.3.3 Model (II): $M/E_k/1$

In the previous model, putting $k_1 = 1$ gives the system $M/E_k/1$. If we write p_n as the steady-state probability of n phases in the system, by writing the steady-state equations to derive the moment generating function, we obtain

$$p_n = (1 - k\rho) \sum_n \rho^m \cdot (-1)^i \binom{m}{i} \binom{m + j - 1}{j},$$

where ρ is the traffic intensity, and the summation is taken over n such that $n = j + ik + m$ (cf. Saaty). From this, the average number of people in the queue may be derived as

$$L = \frac{(k + 1)k\rho^2}{2(1 - k\rho)},$$

and the expected wait W is

$$W = \frac{\rho(1 + k)}{2\mu(1 - k\rho)}$$

where $1/\mu$ is the mean service time.

5.3.4. Model (III): $E_k/M/1$

Here $k_2 = 1$ in the notation of section 5.3.2. To simplify the algebra, Jackson and Nickols wrote the steady-state equation in terms of $n = kn_2 + n_1$, with n_1 and n_2 as before, and derived the moment generating function $P(z)$. Summing over n, the probability p_n of n people in the system is derived as

$$p_o = 1 - \rho$$

$$p_n = \rho(1 - z_o^{-k})z_o^{-(n-1)k} \qquad (n > o) ,$$

where z_o is the unique zero of $P(z)$ outside the unit circle in the complex plane.

Also given are the probabilities q_n that an arriving person will find n people already in the system:

$$q_n = p_n \quad \text{for} \quad k = 1 ,$$

otherwise

$$q_o = 1 - z_o^{-k}$$

$$q_n = (1 - z_o^{-k})z_o^{-(n-1)k} \qquad (n > o).$$

The waiting time distribution is

$$w(t) = z_o^{-k} \lambda (z_o - 1)\exp\left[-\lambda(z_o - 1)t\right] ,$$

where k/λ is the mean of the inter-arrival distribution $dG(t)$.

5.3.5. Model (IV): D/M/1

In the above model, we let k and $\lambda \to \infty$ in such a way that $k/\lambda = 1$. The inter-arrival time distribution then degenerates into the regular distribution of interval 1:-

$$G(t) = \begin{cases} 0 & \text{for } t < 1 \\ 1 & \text{for } t \geq 1 \end{cases}$$

If we let

$$z_o = 1 + \frac{y}{k}$$

and

$$\rho y = 1 - e^{-y} + O(k^{-1}) ,$$

then

$$z_o = 1 + \frac{y_o}{k} + O(k^{-2})$$

where y_o is the real positive root of the equation

$$\rho = \frac{1 - e^{-y}}{y}$$

We may substitute into p_n for the previous model, and obtain

$$p_o = 1 - \rho$$

$$p_n = \rho(1 - e^{-y_o}) e^{-(n-1)y_o} \qquad (n > o)$$

and

$$q_n = \rho y_o (1 - \rho y_o)^n = (1 - e^{-y_o}) e^{-ny_o} \tag{5.5}$$

Also

$$w(t) = e^{-y_o} \sigma(1 - e^{-y_o}) \exp\left[- \sigma(1 - e^{-y_o}) t \right] \tag{5.6}$$

where $1/\sigma$ is the mean service time.

5.3.6. Models (V): $D/E_k/1$ and (VI): $D_m/E_k/1$

For the system $E_p/E_k/1$, we needed the evaluation of the set $\{z_i ; i = 1,2,....k\}$, the roots of

$$\left(\frac{\sigma + s}{\sigma}\right)^k \left(\frac{a - s}{s}\right)^p - 1 = 0 \tag{5.7}$$

with $Re(z_i) < 0$. If we let $p \to \infty$ and $a \to \infty$ in such a way that $p/a = 1$ as before, we have D for the input process, and (5.7) becomes

$$\left(\frac{\sigma + s}{\sigma}\right)^k e^{-s} - 1 = 0 .$$

Alternatively one could attempt an analysis similar to that of Jackson and Nickols and let $p \to \infty$ as before. However, the algebra is difficult and probably not worth pursuing in this context. Similar considerations also apply to the solution of $D_m/E_k/1$.

Bulk queues occur when either arrivals or service no longer take place in single units; instead, a batch of people arrive or are served together. In general the analysis of such systems is more difficult, and in particular the literature concerning bulk arrival queues is quite limited. Gaver (1959) has studied the system in which a variable number of arrivals occurs at each batch, and the intervals between batches are distributed exponentially. The mean waiting time when batch sizes are constant is

$$\frac{\rho}{2(1-\rho)} \cdot \frac{\text{Var}(t)}{E(t)} ,$$

where t is the service time. Thus for an E_k service distribution, this is

$$\frac{\rho}{2(1-\rho)} \cdot \frac{b}{k} .$$

For low values of ρ we would expect this to be approximately true for the system $D_m/E_k/1$. However a generalisation of Gaver's method to the Erlangian inter-arrival time distribution appears difficult.

5.4. Model (VII): $(M + D_m)/M/1$

This model contains the most general input process to be considered. It represents a stream of arrivals composed of two types, in batches of m at regular intervals, and in single units at random. The two classes are represented in some ratio r. The greatest mathematical difficulty in the solution of this system arises because the intervals between arrivals are no longer independent. Points in time just after regular bulk arrivals are points of regeneration, and the numbers in the system at such instants follow a Markov chain. However when we consider all arrival instants, successive intervals between them are correlated.

5.4.1. Structure of the Input Process

There are four types of inter-arrival interval:-

A) Between two regular arrivals

B) Between two random arrivals

C) Beginning with a random arrival and ending with a regular arrival

D) Beginning with a regular arrival and ending with a random arrival

If the sequence of intervals came from some system where all intervals were independent, we would expect the probability of a specific interval having a certain length to be unaffected by the previous history of the input or

the previous sequence of interval types. This is clearly not the case with our present system. For example, an A interval cannot follow a B interval, C cannot follow A, and so on. Consider also the following example: suppose we have a B interval of length x; then a C interval of length $(T_o - x)$ immediately following this has finite probability, where T_o is the constant interval between regular arrivals. B and C type intervals thus have a negative first serial correlation.

It was decided to investigate whether these departures from independence were sufficient to warrant solving the system (M + D)/M/1. If the dependence between arrival intervals was weak, it might be possible to assume complete independence and use the known results from the systems M/M/1 or D/M/1; we might be able to construct a system of the mixed input type by decomposition into two simpler queues of these types, and use their known properties and results, weighted appropriately. We first consider the overall distribution of inter-arrival intervals, and then investigate the structure of the ordered sequence of these intervals.

5.4.2. Distribution of Inter-Arrival Intervals

We denote the stream of random arrivals by S_1, having a rate α_1, and the regular stream by S_2 with arrival rate α_2. We choose at random any arrival from the total input. With probability $\alpha_1/(\alpha_1 + \alpha_2)$ this arrival is from S_1, and with probability $\alpha_2/(\alpha_1 + \alpha_2)$ from S_2. We now consider the distribution of the time t to the next arrival.

(i) Chosen arrival is from S_1

We suppose that our selection from the total input was a random (or unscheduled) arrival; without loss of generality we call the time of this arrival $t = 0$. Then the time to the next regular arrival is uniformly distributed over the time interval $(0, 1/\alpha_2)$ or $(0, T_o)$, where T_o is the regular arrival interval. If this regular arrival is the next arrival in the total input (giving a type C interval), we must have no arrivals from S_1 in $(0, t)$, where t is the next regular arrival time; this event has probability $e^{-\alpha_1 t}$.

However if a random arrival is the next in the joint input (a type B interval), there must be no regulars in $(0, t)$, where t is now the random arrival time; this occurs with probability $\alpha_2 t$. The distribution of the time to the next arrival from S_1 is $\alpha_1 e^{-\alpha_1 t}$.

Combining these results, the cumulative distribution function of the time t to the next arrival in the whole input stream is

$$1 - e^{-\alpha_1 t} . (1 - \alpha_2 t) \quad \text{for } 0 \leqslant t < T_o$$

and the corresponding density is

$$e^{-\alpha_1 t} \{\alpha_1 + \alpha_2 - \alpha_1 \alpha_2 t\}. \qquad (5.8)$$

(ii) Chosen arrival is from S_2

We now suppose that the random selection from the total input was a regular (scheduled) arrival. The distribution of the time until the next arrival is continuous if the next arrival is from S_1 (a type D interval), but there is a probability saltus at $t = T_o$, the time of the next S_2 arrival. The continuous part of the distribution is that of the S_1 stream, i.e. $\alpha_1 e^{-\alpha_1 t}$. If the next arrival is regular from S_2 (a type A interval), then there is no S_1 arrival in the interval $[0, 1/\alpha_2)$; this event has probability

$$1 - \int_0^{1/\alpha_2} \alpha_1 e^{-\alpha_1 t}\, dt = e^{-\alpha_1/\alpha_2} .$$

So the probability density in this case is

$$\alpha_1 e^{-\alpha_1 t} \qquad \text{for} \quad 0 \leq t < 1/\alpha_2$$

and

$$\text{pr}(t = 1/\alpha_2) = e^{-\alpha_1/\alpha_2} \tag{5.9}$$

Combining the two conditional densities (5.8) and (5.9), the probability density for inter-arrival intervals for the whole input stream is

$$\left(\frac{\alpha_1}{\alpha_1 + \alpha_2}\right)(\alpha_1 + \alpha_2 - \alpha_1\,\alpha_2 t)e^{-\alpha_1 t} + \left(\frac{\alpha_2}{\alpha_1 + \alpha_2}\right)\alpha_1 e^{-\alpha_1 t}$$

$$= \left(\frac{\alpha_1}{\alpha_1 + \alpha_2}\right)(\alpha_1 + 2\alpha_2 - \alpha_1\,\alpha_2 t)e^{-\alpha_1 t} \qquad \text{for} \quad 0 \leq t < 1/\alpha_2$$

and a saltus of

$$\left(\frac{\alpha_2}{\alpha_1 + \alpha_2}\right)e^{-\alpha_1/\alpha_2} \qquad \text{at} \quad t = 1/\alpha_2 \qquad (5.10)$$

5.4.3. Structure of Ordered Sequence of Intervals

The density (5.10) represents the overall frequencies of different inter-arrival times, but successive intervals are not independent variables having this distribution. It was decided to consider the behaviour of the first serial correlation coefficient of the arrival interval sequence, that is, the correlation between successive pairs of arrival intervals. For an independent input, we would expect all the serial correlations to be zero on the whole sequence, and also for any subsequence.

To obtain estimates of the correlation, an arrival stream was simulated for various values of r , the ratio of regular to random arrivals, in the region of interest, 0.2 to 5.0, and a few additional extreme values for further information. In order to assess the importance of the B - C correlation, the coefficients are evaluated for the sequence of intervals excluding any pairs of intervals involving an A-type interval, in addition to being evaluated for the whole sequence. The reason for this is clearer on consideration of Figure 5.1 which shows the regions in which interval pairs lie, and the corresponding probability density.

The A-A pairs give a point probability at the point (T_o,T_o); A-B and C-A give line probabilities on $y=T_o$ and $x=T_o$; some of the B-C pairs give a line probability on $y=T_o-x$. The other possible pairs give a probability density over the remaining regions. The density is symmetric about $y = x$, but tends to become higher as the point (0,0) is approached. Although the absolute magnitudes of these point, line and area probability densities are of only secondary importance, it is instructive to examine the general structure of the input sequence and the relative magnitudes of the resultant probabilities in order to account for the behaviour of the correlation coefficients.

Figure 5.2 shows a typical distribution of intervals for low values of r. This implies we have an input of predominantly random arrivals, and most intervals fall in the south-west corner of the available region. A few B-C intervals occur on the line $y=T_o-x$, and even fewer pairs occur on the boundaries $y=T_o$ and $x=T_o$. Figure 5.3 shows the typical behaviour for rather larger values of r. More points occur on the boundaries and the B-C line, and also A-A pairs begin to occur. The overall area densities are reduced. Figure 5.4 shows a sample behaviour for a high value of r, which implies an input-process consisting almost entirely of regular arrivals. A-A pairs now occur with high frequency. Of the rest, many are on the boundaries (implying one of the intervals of the pair was A-type), some are on the B-C line, and very few remain in the area density region.

For most values of r, the system was simulated 20 times with 2,000 arrivals each. In several of the higher and outlying values of r, 4,000 arrivals were

FIGURE 5.1 LOCATION OF INTER-ARRIVAL INTERVAL PAIR TYPES WITH INPUT MECHANISM ($M + D_m$)

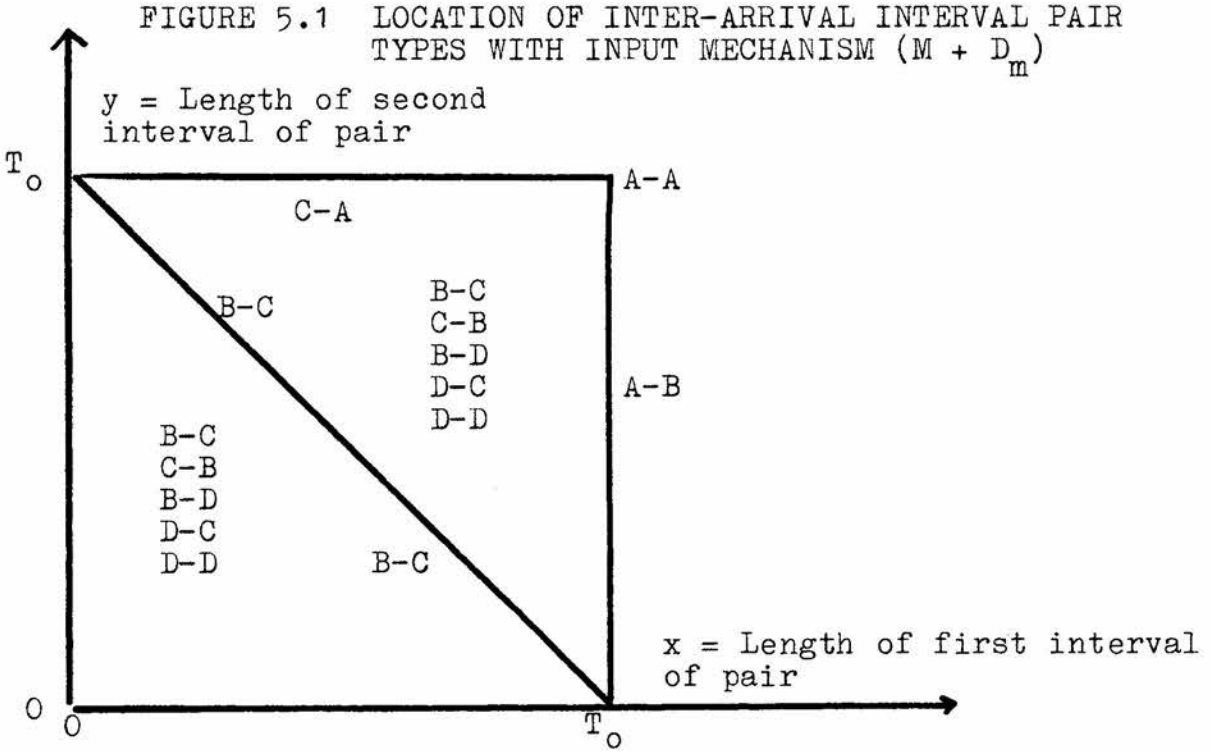


FIGURE 5.2 SAMPLE DISTRIBUTION OF PAIR TYPES FOR A LOW VALUE OF r

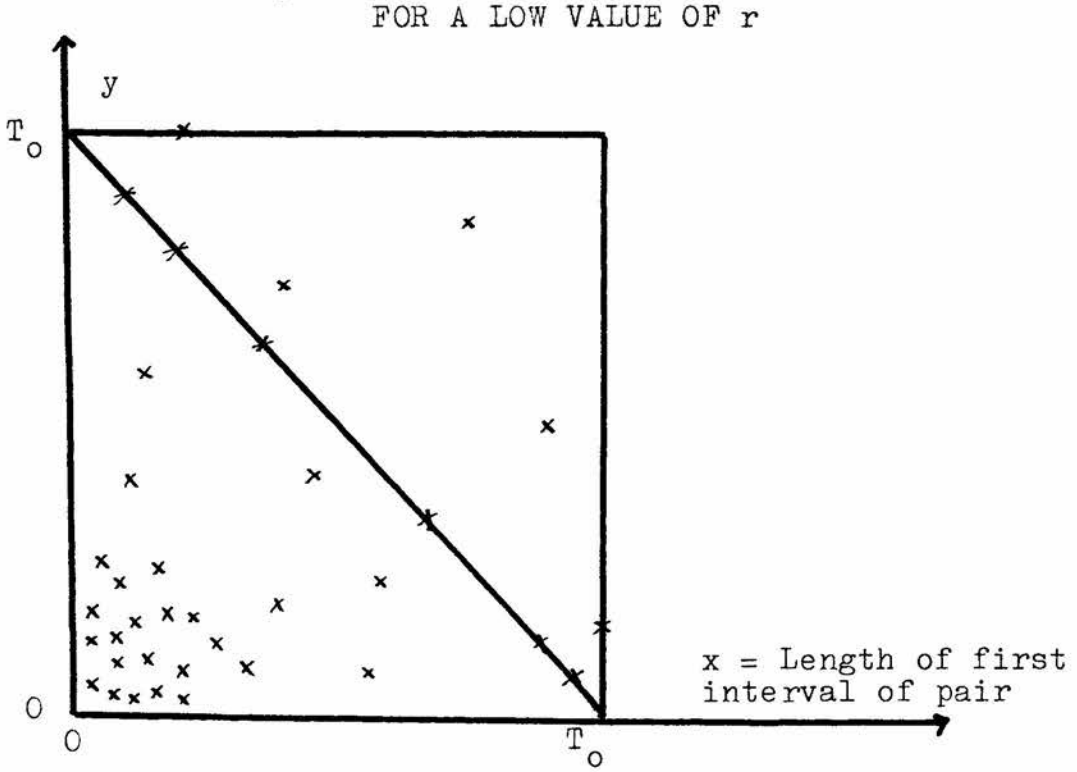


FIGURE 5.3 SAMPLE DISTRIBUTION OF PAIR TYPES FOR AN INTERMEDIATE VALUE OF r

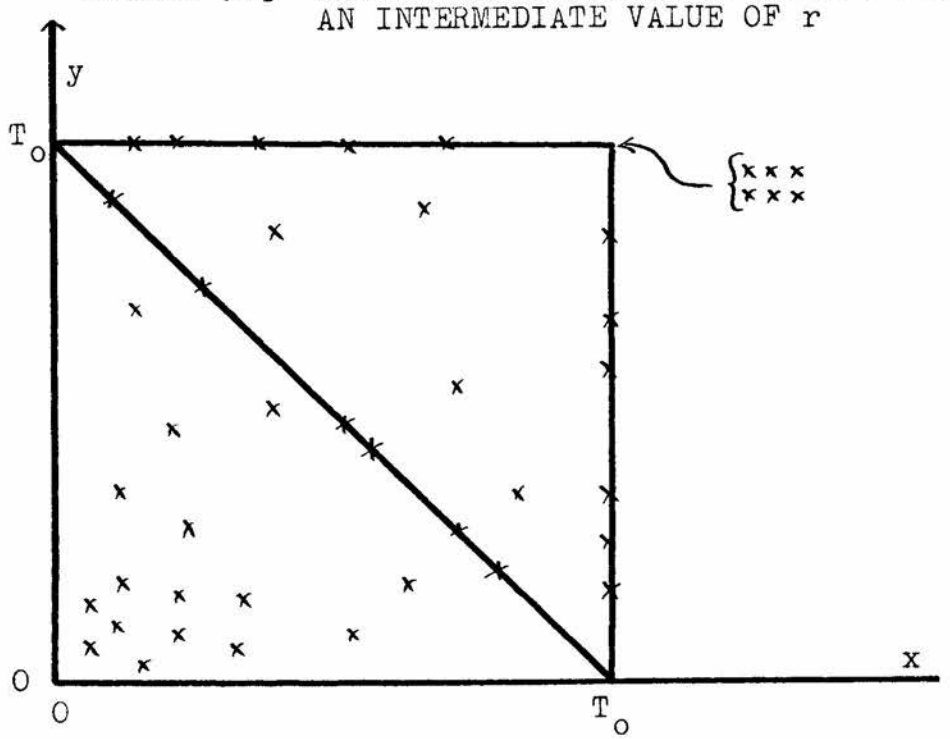
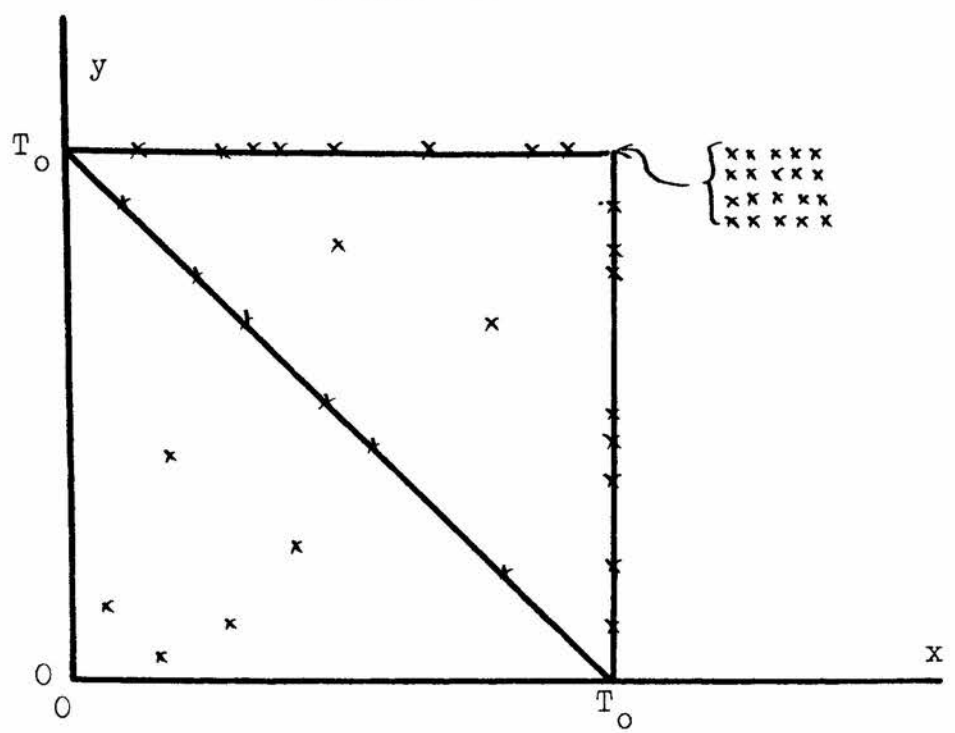


FIGURE 5.4 SAMPLE DISTRIBUTION OF PAIR TYPES FOR A HIGH VALUE OF r



generated in each run, and for some of the lower r values 1,500 arrivals were used. The results are given in Table 5.1 and Figure 5.5.

We define c_1 as the first serial correlation coefficient for the whole sequence of interval pairs, and c_2 as the first serial correlation coefficient for the whole sequence except pairs involving at least one A-type. It may be seen that in the range $0.2 < r < 5.0$, c_1 is fairly small. It is only when c_2 is evaluated that it becomes clear that the relevant subsequence contains elements which are far from being independent (as indicated by this means). In fact c_2 decreases almost exponentially in the range $0.35 < r < 5.00$, and there is already an appreciable correlation at the value $r = 1.0$.

Limit behaviour of c_1 and c_2

(i) As $r \to \infty$, $c_2 \to -1$. All points, except those involving A-pairs, tend to occur on the B-C line. Thus for a high value of r, a typical sequence might be AAA...ABCAA.... . In other words, after a random arrival (a rare event), the most likely next arrival is regular.

(ii) As $r \to \infty$, $c_1 \to 1$. The great majority of pairs occur at the A-A corner. The convergence seems much slower than that of c_2.

(iii) As $r \to 0$, $c_1 \to 0$ and $c_2 \to 0$. Almost no regular arrivals occur, giving a typical sequence of DDD...DCBDD... . The input, which is almost entirely random, has very small correlation.

For low values of r, B-C intervals have greater frequency than A-pairs, and c_1 is negative until increasing r also increases the relative frequency of A-pairs.

TABLE 5.1 FIRST SERIAL CORRELATION OF INTER-ARRIVAL INTERVALS WITH THE ARRIVAL MECHANISM ($M + D_m$)

r	Number of arrivals per simulation run	c_1	c_2
0.10	1500	- 0.0032	- 0.0032
0.20	2000	- 0.0368	- 0.0405
0.35	1500	- 0.0495	- 0.0848
0.50	2000	- 0.0537	- 0.1483
0.75	1500	- 0.0224	- 0.2331
1.00	2000	+ 0.0163	- 0.2993
1.50	4000	+ 0.0588	- 0.4134
2.00	2000	+ 0.0909	- 0.4846
5.00	2000	+ 0.1686	- 0.7298
10.00	4000	+ 0.2132	- 0.8377
100.00	4000	+ 0.2492	- 0.9954

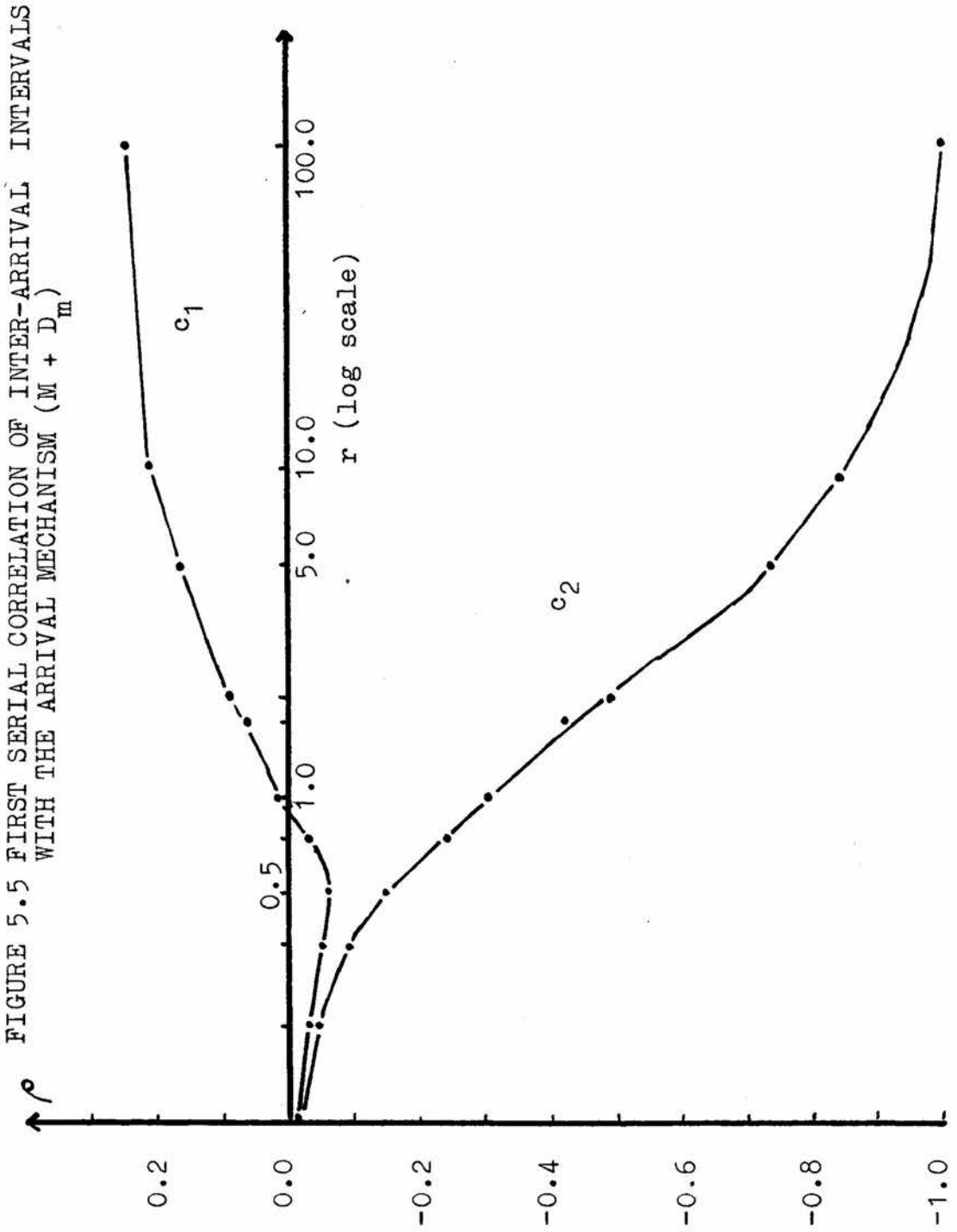


FIGURE 5.5 FIRST SERIAL CORRELATION OF INTER-ARRIVAL INTERVALS WITH THE ARRIVAL MECHANISM ($M + D_m$)

We have seen that the dependence between arrival intervals is not weak when a subsequence of the intervals is considered. The serial correlation is significantly different from zero in the larger part of the range of interest of r. Consequently it was felt that an approximation to the system (M + D)/M/1 by an equivalent system M/M/1 or D/M/1 might be inadequate for the present purpose. However there was still the possibility that numerical results might show that these latter two models would give comparable results, or possibly some weighted average of the two solutions would yield an approximation to (M + D)/M/1.

5.4.4 Solution of the System. $(M + D_m)/M/1$

Without loss of generality, we take the interval between regular arrivals as the time unit. For a general interval, a simple scale transformation of the results is required. We denote the constant batch size by m, the rate of single random arrivals by α, and the service rate by σ.

5.4.5 Steady-state Solution

Firstly we note that the instants of regular arrivals are regeneration points, and that the numbers in the system just after such times from a Markov chain. We also note that in the intervals between regular arrivals we have fragmentary realisations of the simple queue M/M/1.

Suppose at some instant that the system is in state i ; then the probab-

ility that there are j in the system after a time t we denote by $p_{ij}(t)$, the general time-dependent transition probability. For the simple queue this is known explicitly as

$$p_{ij}(t) = \rho^{(j-i)/2} \cdot e^{-(\alpha+\sigma)t} \left\{ I_{j-i} + \rho^{-\frac{1}{2}} \cdot I_{i+j+1} + (1-\rho)\rho^{-\frac{1}{2}} \sum_{k=1}^{\infty} I_{i+j+1+k} \cdot \rho^{-k/2} \right\} \qquad (5.11),$$

where $\rho = \alpha/\sigma$ is the traffic intensity, and $I_n = I_n(2t\sqrt{\alpha\sigma})$ is the modified Bessel function of integer order n with argument $2t\sqrt{\alpha\sigma}$.

Let

$$\underline{\pi}^T = (\pi_m, \pi_{m+1}, \pi_{m+2}, \ldots\ldots)$$

be the probability distribution of the number in the system at instants immediately after regular arrival times. Also we let

$$p(i,j) = p_{ij}(1) \quad ,$$

the transition probability for the particular time of the appointment interval. If there are j in the system just before any given regular arrival time, there are j + m just after it. As the system is in equilibrium, we must have

$$\pi_{j+m} = \sum_{i=m}^{\infty} \pi_i \, p(i,j) \qquad j = 0,1,2, \ldots. \qquad (5.12).$$

If we define $\underline{P}$ as the infinite matrix

$$\begin{bmatrix} p(m,0) & ; & p(m,1) & ; & p(m,2) & ; & \cdots \\ p(m+1,0) & ; & p(m+1,1) & ; & p(m+1,2) & ; & \cdots \\ p(m+2,0) & ; & p(m+2,1) & ; & p(m+2,2) & ; & \cdots \\ \vdots & & \vdots & & \vdots & & \end{bmatrix}$$

then equations (5.12) can be written as

$$\underline{\pi}^T = \underline{\pi}^T \cdot \underline{P}$$

or

$$\underline{\pi}^T \cdot (\underline{I} - \underline{P}) = \underline{0} \qquad (5.13)$$

where $\underline{I}$ is the infinite identity matrix, and $\underline{0}$ is the infinite zero vector. We also require the normalising equation

$$\sum_{i=m}^{\infty} \pi_i = 1 \qquad (5.14)$$

For the purpose of numerical solution of the infinite set of equations (5.13) and (5.14), we define a truncated approximation vector

$$\underline{\pi'}^T = (\pi'_m, \pi'_{m+1}, \pi'_{m+2}, \cdots)$$

to $\underline{\pi}^T$ which satisfies the finite set of equations:-

$$
\begin{aligned}
\pi'_m &= \pi'_m p(m,0) + \pi'_{m+1} p(m+1,0) + \ldots + \pi'_{m+N-1} p(m+N-1,0) \\
\pi'_{m+1} &= \pi'_m p(m,1) + \pi'_{m+1} p(m+1,1) + \ldots + \pi'_{m+N-1} p(m+N-1,1) \\
&\vdots \qquad\qquad \vdots \qquad\qquad \vdots \qquad\qquad\qquad\qquad \vdots \\
\pi'_{m+N-2} &= \pi'_m p(m,N-2) + \pi'_{m+1} p(m+1,N-2) + \ldots \pi'_{m+N-1} p(m+N-1, \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad N-2)
\end{aligned}
\tag{5.15}
$$

and

$$
\sum_{i=m}^{m+N-1} \pi'_i = 1 \tag{5.16}
$$

It is clear that as $N \to \infty$, $\pi'_i \to \pi_i$ for all $i \geq m$; also if a true equilibrium $\underline{\pi}$ exists, then $\pi_i \to 0$ as $i \to \infty$. We wish to choose N sufficiently large that $\underline{\pi}'$ is in some sense a sufficiently good approximation to the appropriate subvector

$$
(\pi_m, \pi_{m+1}, \cdots \quad \pi_{m+N-1})
$$

of $\underline{\pi}^T$, and also that

$$
\sum_{i=m+N}^{\infty} \pi_i < \varepsilon \quad ,
$$

where ε is some small specified quantity.

Further difficulties arise in the numerical solution of the truncated system of equations (5.15) and (5.16) when we come to evaluate the coefficients p(i,j). Each one involves an infinite sum of weighted Bessel functions (see equation (5.11)), and we must choose an upper index for the sum in an approximation. Section 5.5 considers the evaluation of the p(i,j) elements of $\underline{P}$, and section 5.6 considers the dimension N of the set of equations (5.15) and (5.16).

5.5. Evaluation of the Elements p(i,j)

We denote by $\underline{P}'$ the truncated matrix corresponding to $\underline{P}$ in equations (5.15); i.e.

$$\underline{P}' = \begin{bmatrix} p(m,0) & ; & p(m,1) & ; & \cdots & p(m,N-2) \\ p(m+1,0) & ; & p(m+1,1) & ; & \cdots & p(m+1,N-2) \\ \vdots & & \vdots & & & \vdots \\ p(m+N-1,0); & & p(m+N-1,1); & & \cdots & p(m+N-1,N-2) \end{bmatrix}$$

The coefficient p(i,j) involves the infinite sum

$$S = \sum_{k=1}^{\infty} \rho^{-k/2} \cdot I_{i+j+1+k}(2\sqrt{\alpha\sigma})$$

(taking t = 1). For the numerical evaluation of p(i,j), we wish to establish some value N' such that

$$\sum_{\nu=N'}^{\infty} \rho^{-\frac{\nu}{2}+\frac{c}{2}} \cdot I_{\nu}(x) < \varepsilon ,$$

for given values c, ρ, x and ε.

5.5.1. Order of Evaluation

For a particular value n of N, we have n^2 elements p(i,j) to evaluate, each requiring the use of a fairly large number of Bessel functions. A substantial proportion of the computing time was saved by noting from the form of p(i,j) that if (i - j) is constant, then p(i,j) tends to a limit as i tends to infinity. To utilise this fact profitably, the elements of $\underline{P}'$ were evaluated in row order until p(i,0) was less than some specified small quantity ε_1 ; the (i + 1)th. row was then substituted from row i by the relations

$$p(i + 1,0) = 0$$

and

$$p(i + 1,j) = p(i,j - 1) \quad j = 1,2,\ldots N - 1.$$

The process was continued up to row (N - 1). In practice for $\varepsilon_1 = 10^{-6}$ and typical values of ρ and x, usually only the first four or five rows were evaluated by individual elements before p(i,0) had become sufficiently small. The extra computing time involved in filling the elements of $\underline{P}'$ in row order, rather then the more usual and faster column order, was greatly outweighed by the saving in the evaluation of elements which were almost identical.

5.5.2 An Approximation to S

Over the ranges of the parameters we are considering, the sum S varies greatly in its rate of convergence. It would not

only be highly inefficient to take some common value N' for all elements p(i,j), but also this would create additional practical difficulties in a computer evaluation, as is demonstrated below.

The sequence $\{I_k(x)\}$ tends to the limit zero more rapidly for smaller x. In our example,

$$x = 2\sqrt{\alpha\sigma} = \sqrt{r\rho_s}^{2m} ,$$

where $\rho_s = \rho - m/\sigma$, the traffic intensity related to the random inputs; ρ_s is also the appropriate intensity for the simple queue realisations between regular arrival times. In the X-Ray situation, m is never higher than 5 in practice, and r may range between 0.2 and 5.0. Clinics exist where the input is almost all random or all regular (giving r = 0 and ∞) but these give models (I): M/M/1 and (IV): D/M/1 respectively, with known solutions. Where the input is normally a mixture of the two types, the range (0.2, 5.0) of r will cover most examples. We consider ρ in the range 0.4 to 0.9; values of ρ less than 0.4 will represent clinics where there is little congestion anyway; values above 0.9 and close to 1.0 will show few characteristics of equilibrium behaviour and the model will be a very poor representation of the real system. We thus consider the values

$$m = 1,2,3,4,5$$

$$r = 0.2,\ 0.5,\ 1.0,\ 1.5,\ 2.0,\ 5.0$$

and ρ .4, 0.5, 0.6, 0.7, 0.8 and 0.9.

We note that we have more rapid convergence of S for small m, large r and large ρ_s. When the parameter triple (m,r,ρ) is (1,5.0,0.9), we have the minimum value of x as 1.03 (over the parameter values considered), and the maximum is

$x = 86.6$ when $(m,r,\rho) = (5, 0.2, 0.4)$.

5.5.3 Evaluation of $I_\nu(x)$

Table 5.2 of some specific values of Bessel functions shows great variations in both the magnitudes of individual functions and the rates of convergence to zero of the sequence $\{I_\nu(x)\}$ as $\nu \to \infty$ for different values of x.

The computer used in this work was able to store real numbers in the range of magnitudes 16^{-65} to 16^{63} or approximately 10^{-78} to 10^{75}. An attempt to store a number less than 10^{-78} gave an error known as underflow; the usual corrective action by the computer was to replace the number being stored by zero and to continue the program. However only a limited number of underflows were allowed in each step of the job before the program was automatically terminated. With such a large number of potential underflows in the evaluation of $\underline{P}'$ for each parameter combination, it was clearly desirable to eliminate them if possible. This was not possible by using logarithms, as a sum of small functions was required; also the method of successive evaluation of the Bessel functions made the use of scaling or logarithms difficult in practice.

An I.B.M. supplied scientific subroutine package was used to calculate the Bessel functions. Firstly a value of $I_o(x)$ was evaluated using a polynomial expansion. Then using the recurrence relation

$$I_{k+1}(x) + \frac{2k}{x} I_k(x) - I_{k-1}(x) = 0$$

TABLE 5.2 LOGARITHMS (BASE 10) OF MODIFIED BESSEL FUNCTIONS OF INTEGER ORDER

x \ N	1	5	10	20	50	100
1.0	- 0.3	- 3.6	- 9.6	- 24.4	- 79.5	-188.1
5.0	1.4	0.3	- 2.3	- 10.3	- 44.5	-118.1
10.0	3.4	2.9	1.3	- 3.9	- 29.3	- 88.0
20.0	7.6	7.4	6.5	3.5	- 13.6	- 57.5
50.0	20.5	20.3	20.0	18.7	10.3	- 15.6
80.0	33.4	33.3	33.1	32.3	26.8	8.7

and also

$$G_{k+1}(x) = I_k(x)/I_{k-1}(x)$$

$$= \cfrac{1}{\cfrac{2(k+1)}{x}} + \cfrac{1}{\cfrac{2(k+2)}{x}} + \ldots \qquad \text{(continued fraction)}$$

the values of $I_n(x)$ for $n = 1,2,\ldots N'$ are found using the given value of $I_o(x)$. It seemed much more efficient to determine the value N' dependent on x, and evaluate all the functions required by one use of the subroutine. An alternative would be to call the subroutine, check the size of the highest order function evaluated (to guard against underflow), and recall the subroutine to generate the next function. This would require a very large number of such checks and the method would become very inefficient in the use of computer time. It was decided to establish upper bounds for N' and evaluate all the functions for a particular parameter triple (m,r,ρ) by one calling of the subroutine.

5.5.4 Bounds for N'

We wish to determine some value of N' which satisfies

$$\sum_{k=1}^{\infty} \rho^{-k/2} I_{i+j+1+k}(2\sqrt{\alpha\sigma}) - \sum_{k=1}^{N'-1} \rho^{-k/2} . I_{i+j+1+k}(2\sqrt{\alpha\sigma}) < \varepsilon$$

or

$$\sum_{\nu=N'}^{\infty} \rho^{-\frac{\nu}{2} + \frac{c}{2}} . I_\nu(x) < \varepsilon,$$

where ρ here refers to the simple queue traffic intensity ρ_s, and $c = i + j + 1$.

We may expand $I_\nu(x)$ as

$$I_\nu(x) = \left(\frac{x}{2}\right)^\nu \sum_{n=0}^{\infty} \frac{\left(\frac{x^2}{4}\right)^n}{n!\,(\nu + n)!}$$

Thus

$$S_1 = \sum_{\nu = N'}^{\infty} \rho^{-\nu/2} . I_\nu(x)$$

$$= \sum_{\nu = N'}^{\infty} \rho^{-\nu/2} . \left(\frac{x}{2}\right)^\nu \sum_{n=0}^{\infty} \frac{\left(\frac{x^2}{4}\right)^n}{n!\,(\nu + n)!}$$

$$= \sum_{\nu = N'}^{\infty} \sum_{n=0}^{\infty} \frac{(x/2)^{2n + \nu} . \rho^{-\nu/2}}{n!\,(\nu + n)!}$$

$$= \sum_{\nu = N'}^{\infty} \rho^{-\nu/2} \sum_{n=0}^{\infty} \frac{(x/2)^n}{n!} \cdot \frac{(x/2)^{n+\nu}}{(n + \nu)!}$$

$$\leqslant \sum_{\nu = N'}^{\infty} \rho^{-\nu/2} \sum_{n=0}^{\infty} \frac{(x/2)^n}{n!} \cdot \frac{(x/2)^n}{n!} \cdot \frac{(x/2)^\nu}{\nu!}$$

$$\leq \sum_{\nu = N'}^{\infty} \rho^{-\nu/2} \left(\sum_{n=0}^{\infty} \frac{(x/2)^n}{n!} \right)^2 \frac{(x/2)^\nu}{\nu!}$$

$$= (e^{x/2})^2 \sum_{\nu = N'}^{\infty} \frac{(x/2\sqrt{\rho})^\nu}{\nu!}$$

$$= e^x \left\{ \frac{(x/2\sqrt{\rho})^{N'}}{N'!} + \frac{(x/2\sqrt{\rho})^{N'+1}}{(N'+1)!} + \ldots \right\}$$

$$\leq e^x \left\{ \frac{(x/2\sqrt{\rho})^{N'}}{N'!} + \frac{(x/2\sqrt{\rho})^{N'}}{N'!} \cdot \frac{(x/2\sqrt{\rho})}{N'} + \frac{(x/2\sqrt{\rho})^{N'}}{N'!} \cdot \frac{(x/2\sqrt{\rho})^2}{N'^2} + \ldots\ldots \right\} \qquad (5.17)$$

If we assume N_1 is sufficiently large that $N_1 > x/2\sqrt{\rho}$, then the terms in the brackets of (5.17) form a convergent geometric series.

Thus

$$S_1 \leq e^x \cdot \frac{(x/2\sqrt{\rho})^{N'}}{n'!} \cdot \frac{1}{1 - x/2N'\sqrt{\rho}} \qquad (5.18)$$

When we include the term in $\rho^{c/2}$, where $c = i + j + 1$, this becomes with the previous notation ρ_s,

$$S_1 \leq \frac{e^x \cdot x^{N'} \cdot \rho_s^{(i+j+1)/2}}{(N'-1)!\,(2\sqrt{\rho_s})^{N'-1}\cdot(2N'\sqrt{\rho_s} - x)} \tag{5.19}$$

In (5.18) we now write $y = x/2\sqrt{\rho}$. Then

$$S_1 \leq e^x \cdot \frac{y^{N'}}{N'!} \cdot \frac{1}{1 - y/N'}$$

$$= e^x \cdot \frac{y^{N'}}{(N' - y)} \cdot \frac{1}{(N'-1)!}$$

If we now assume $N' - y > y$, or $N' > 2y = x/\sqrt{\rho}$, then we have

$$S_1 \leq \frac{e^x \cdot x^{N'-1} \cdot \rho_s^{(i+j+1)/2}}{(N'-1)!\,(2\sqrt{\rho_s})^{N'-1}} \tag{5.20}$$

which is a smaller upper bound for S_1 than that of (5.19).

Our object now is to find the smallest values of N', N* say, such that the quantities on the right-hand sides of (5.19) and (5.20) are less than some small number ϵ, the order of accuracy of the approximation; ϵ was taken here as 10^{-4}. After some trial and error to find a few particular values of N*, it appeared that the condition $N^* > x/\sqrt{\rho}$, necessary for the upper bound given by (5.20), was frequently not satisfied, whereas the N* given by

(5.19) did satisfy the relevant condition $N^* > x/2\sqrt{\rho}$ for almost all values of the parameters m, r and ρ. Therefore it was decided to define the upper bound N' as $\text{Max}(N^*, N_1)$, where N_1 is the smallest integer such that $N_1 > x/2\sqrt{\rho}$, and N^* is the smallest integer such that

$$\frac{e^x \cdot x^{N^*}}{(N^* - 1)!} \cdot \frac{\rho_s^{(i + j + 1)/2}}{(2\sqrt{\rho_s})^{N^* - 1}} \cdot \frac{1}{(2N^*\sqrt{\rho_s} - x)} < \varepsilon \qquad (5.21)$$

5.5.5 Numerical Evaluation of N^*

For each value of the triple (m,r,ρ), we must evaluate $x = 2\sqrt{\alpha\sigma}$, where α is the random arrival rate and σ is the service rate of all customers. We have

r = Rate of regular arrivals/Rate of random arrivals

$= m/\alpha$

$\therefore \alpha = m/r$.

Also ρ = Total rate of arrivals/service rate

$= (\alpha + m)/\sigma$

$\therefore \sigma = (\alpha + m)/\rho$.

Thus

$$x = 2\left(\frac{m(\alpha + m)}{r\rho}\right)^{\frac{1}{2}}$$

Now ρ_s = Total intensity - regular arrivals intensity

$$= \rho - m/\sigma$$

$$= \rho/(1 + r) \qquad (5.22)$$

after substituting for σ and rearranging. Hence

$$x = 2\left(\frac{m(\alpha + m)}{r\rho_s(1 + r)}\right)^{\frac{1}{2}}$$

$$= \frac{2m}{r\sqrt{\rho_s}} \qquad (5.23)$$

substituting for α. For given values of m, r and ρ, the Bessel argument x was evaluated using expressions (5.22) and (5.23). Tables 5.3 and 5.4 give values of ρ_s and x/m for all the values of ρ and r considered.

For m = 1, exact values of N^* were found, and there are given in Table 5.5. For m = 2,3,4 and 5, upper bounds on N^* were found partly by using the original trial and error values, and partly by noting that N^* is monotonic decreasing with increasing r and ρ. The results are shown in Table 5.6.

Having established the values N^* to be used, it was necessary to check that the magnitude of $I_{N^*}(x)$ was not less than 10^{-78} for storage in the computer. For a particular N^* and appropriate z, the following uniformly asymptotic expansion of $I_\nu(\nu z)$, with large values of ν, was used:

TABLE 5.3 VALUES OF $\rho_s = \rho/(1 + r)$

r / ρ	0.2	0.5	1.0	1.5	2.0	5.0
0.4	0.333	0.267	0.200	0.160	0.133	0.067
0.5	0.417	0.333	0.250	0.200	0.167	0.083
0.6	0.500	0.400	0.300	0.240	0.200	0.100
0.7	0.582	0.467	0.350	0.280	0.233	0.117
0.8	0.667	0.533	0.400	0.320	0.267	0.133
0.9	0.750	0.600	0.450	0.360	0.300	0.150

TABLE 5.4 VALUES of x/m

r / ρ	0.2	0.5	1.0	1.5	2.0	5.0
0.4	17.32	7.74	4.42	3.33	2.74	1.55
0.5	15.49	6.93	4.00	2.98	2.34	1.39
0.6	14.15	6.32	3.65	2.72	2.24	1.27
0.7	13.11	5.86	3.38	2.53	2.07	1.17
0.8	12.25	5.48	3.16	2.36	1.94	1.10
0.9	11.54	5.16	2.98	2.22	1.83	1.03

TABLE 5.5 VALUES OF N^* FOR m = 1

r / ρ	0.2	0.5	1.0	1.5	2.0	5.0
0.4	60	33	23	21	19	16
0.5	51	28	20	18	15	14
0.6	45	25	18	16	15	13
0.7	39	22	16	14	13	12
0.8	36	20	15	13	12	11
0.9	32	19	14	12	11	10

TABLE 5.6 N^* FOR m = 2, 3, 4 AND 5

ρ	r / m	0.2	0.5	1.0	1.5	2.0	5.0
0.4	2	125	65	41	38	34	27
	3	170	89	58	50	45	36
	4	230(200)	120	76	63	58	45
	5	290(240)	150	95	85	65	55
0.5	2	100	55	40	35	30	27
	3	150	83	55	50	45	35
	4	190	110	70	65	55	40
	5	240(215)	140	90	80	60	50
0.6	2	83	50	35	30	30	25
	3	130	70	50	45	40	35
	4	170	90	65	60	50	40
	5	210	110	85	75	60	45
0.7	2	75	40	30	28	25	24
	3	120	65	45	40	40	35
	4	160	75	60	60	50	40
	5	200	95	80	70	60	45
0.8	2	70	40	30	25	24	22
	3	105	60	45	40	30	20
	4	135	75	50	40	35	35
	5	165	90	75	60	40	40
0.9	2	60	34	24	20	17	14
	3	88	49	30	28	25	19
	4	120	60	37	33	30	24
	5	155	80	70	60	35	29

$$I_\nu(\nu z) \sim \frac{1}{\sqrt{2\pi\nu}} \cdot \frac{e^{\nu\eta}}{(1+z^2)^{\frac{1}{4}}} \left\{ 1 + \sum_{k=1}^{\infty} \frac{U_k(t)}{\nu^k} \right\}$$

where $t = \frac{1}{\sqrt{1+z^2}}$

and $\eta = \sqrt{1+z^2} + \ell n \left(1 + \frac{z}{\sqrt{1+z^2}}\right) = \frac{1}{t} - \frac{1}{2} \ell n \left(\frac{1+t}{1-t}\right)$. The $U_k(t)$, Debye's functions, are defined by

$$U_o(t) = 1$$

and the recurrence relation

$$U_{k+1}(t) = \tfrac{1}{2}t^2(1-t^2)U_k'(t) + \tfrac{1}{8}\int_o^t (1-5t^2)U_k(t)dt.$$

$$k = 0,1,2,\ldots$$

For this work, where ν is large and z is small giving a value of t close to unity, the functions $\{U_i(t)\}$ form a rapidly converging sequence with limit zero. To establish the approximate size of the Bessel functions, the further approximation

$$I_\nu(\nu t) \sim \frac{1}{\sqrt{2\pi\nu}} \frac{e^{\nu\eta}}{(1+z^2)^{\frac{1}{4}}}$$ was used.

For values of (m,r,ρ) equal to $(4,0.2,0.4)$, $(5,0.2,0.4)$ and $(5,0.2,0.5)$, the values of N^* gave $I_{N^*}(x)$ too small to store in the computer in the normal way. Lower values N_2 of N' were used and are shown bracketed in Table 5.6; these were the largest integers such that $I_{N_2}(x) > 10^{-78}$. Although these values of N' do not satisfy explicitly the criteria of the others, it should be remembered that the sum S may converge to sufficient accuracy a good deal more

rapidly than indicated by the values of N^*; this is because of all the inequalities used in the derivation of (5.21). For large x these inequalities may be great, and the series S may converge sufficiently when only a small proportion of N^* terms is used.

It was also necessary to check that the factor $\rho^{-k/2}$ did not exceed 10^{75} when k took its greatest value; this did not in fact occur. Finally in the computer evaluation of S, the terms were added in descending order of the Bessel functions; this was for improved accuracy in slowly convergent sums.

5.6 Evaluation of N

The second major problem in the numerical solution of equations (5.15) and (5.16) is in the choice of N, the number of terms of the distribution $\underline{\pi}$ to which we wish to make a non-zero approximation. We require that the derived solution $\underline{\pi}'$ is a sufficiently good approximation to the corresponding part

$$(\pi_m, \pi_{m+1}, \cdots \pi_{m+N-1})$$

of the true distribution, and that $\sum_{i=m+N}^{\infty} \pi_i$, the sum of the remaining terms, is small. For the latter, we require that

$$(i) \quad \sum_{i=m+N}^{\infty} \pi_i < \varepsilon_1$$

where ε_1 is a small known quantity. For $\underline{\pi}'$ to be a good approximation we require $\underline{\pi}'$ to satisfy

(ii)a. $|\pi_i - \pi_i'| < \varepsilon_2$ for $i = m, m + 1, \ldots m + N - 1$.

and

(ii)b. $\dfrac{|\pi_i - \pi_i'|}{\pi_i} < c$ for $i = m, m + 1, \ldots m + N - 1$.

where ε_2 and c are some small specified constants.

Condition (i)

Let the solution $\underline{\pi}'$ for a particular choice of N be

$$\underline{\pi}^{(N)} = (\pi_m^{(N)}, \pi_{m+1}^{(N)}, \ldots \pi_{m+N-1}^{(N)}).$$

It is impossible to tell by direct inspection of $\underline{\pi}^{(N)}$ if N satisfies condition (i) as $\underline{\pi}$ is unknown, unless it happens that $\pi_{m+N-1}^{(N)}$ is much larger than ε_1; if this happens we may say that N has been chosen too small. Otherwise we must use a comparison of $\underline{\pi}$ with the known solution when r = o, that is the solution of M/M/1.

The input to the simple queue is random by definition, and we would expect a greater mean waiting time than with an input of regularly spaced arrivals, as in the system D/M/1. In general, higher values of r in the system (M + D)/M/1 give a more deterministic and "spaced" arrival sequence, and a distribution $\underline{\pi}$ results which is more concentrated on n, the number in the system.

For M/M/1, the probability p_n of n being in the system at any given time is

$$p_n = \rho^n (1 - \rho).$$

For comparison with $\underline{\pi}$, we consider

$$\underline{\theta} = (\theta_1, \theta_2, \ldots\ldots \quad)$$

defined as the distribution of the number in the simple system just after any given arrival. All arrivals are independent and at random, and so the set of arrival instants forms a random sample of times from all time instants to which the distribution $\underline{P} = (p_o, p_1, \ldots.)$ applies. We must have therefore

$$\theta_{n+1} = \rho^n(1 - \rho) \qquad n = 0,1,2,\ldots.$$

$m = o$ implies that $r = o$ and in view of the previous remarks this system gives $\underline{\theta}$ as the particular distribution of the general family of distributions $\underline{\pi}$ with greatest dispersion and the longest "tail". Formally, there exists an integer n such that

$$\pi_{i+m} \leqslant \theta_i \quad \text{for} \quad i \geqslant n.$$

Thus

$$\sum_{i=n+m}^{\infty} \pi_i \leqslant \sum_{i=n}^{\infty} \theta_i = \rho^{n-1}.$$

To satisfy condition (i) we must take N such that

$$\rho^{N-1-m} < \varepsilon_1.$$

Conditions (ii)

The method used in the evaluation of the elements of $\underline{P}'$ was to

reduce the system of equations (5.15) to an equivalent system with coefficients a_{ij} which approximate to p(i,j). The new system is of the form:-

$$
\begin{aligned}
\pi_m' &= a_{m,0}\pi_m' + a_{m+1,0}\pi_{m+1}' + \cdots + a_{m+N-1,0}\pi_{m+N-1}' \\
\pi_{m+1}' &= a_{m,1}\pi_m' + a_{m+1,1}\pi_{m+1}' + \cdots + a_{m+N-1,1}\pi_{m+N-1}' \\
&\vdots \\
\pi_s' &= a_{m,s}\pi_m' + a_{m+1,s}\pi_{m+1}' + \cdots + a_{m+N-1,s}\pi_{m+N-1}' \\
\pi_{s+1}' &= 0 + a_{m,s}\pi_{m+1}' + a_{m+1,s}\pi_{m+2}' + \cdots + a_{m+N-2,s}\pi_{m+N-1}' \\
\pi_{s+2}' &= 0 + 0 + a_{m,s}\pi_{m+2}' + \cdots + a_{m+N-2,s}\pi_{m+N-1}' \\
&\vdots \\
\pi_{m+N-2}' &= 0 + 0 + \cdots + a_{m,s}\pi_{2m-N-2-s}' + a_{m+1,s}\pi_{2m+n-1-s}' \\
&\qquad + \cdots + a_{s+1,s}\pi_{m+N-1}'
\end{aligned}
\tag{5.24}
$$

This system gives the greatest relative error in the terms π_i' with largest i.

Suppose now that we solve the model M/M/1 by the above numerical method, taking m = o. As before, a choice of N which gives a satisfactory solution to the simple queue will also give a satisfactory solution to the system (M + D)/M/1 with mixed input of the same total rate. Note that as m = r = 0 in the simple queue solution, we must set the arrival rate α to fix the time scale; we no

longer have regular arrivals to fix the time unit. We take α to be of the same magnitudes as in (M + D)/M/1 to give the same order of rounding errors in the two solutions.

Let the solution $\underline{\theta}$ of the system with m = 0 for a particular choice of N be

$$\underline{\theta}^{(N)} = (\theta_1^{(N)}, \theta_2^{(N)}, \ldots\ldots \theta_N^{(N)}).$$

Then we require

$$|\theta_i - \theta_i^{(N)}| < \varepsilon_2$$

and

$$\frac{|\theta_N - \theta_N^{(N)}|}{\theta_N} < c$$

(as the greatest proportional error occurs in $\theta_N^{(N)}$). The normalising equation (5.16) applied to $\underline{\theta}^{(N)}$ gives $\theta_N^{(N)} > \theta_N$, and so we require

$$\theta_N^{(N)} - \theta_N < c\,\theta_N = c\rho^{N-1}(1-\rho).$$

If we take

$$c\rho^{N-1}(1-\rho) < \varepsilon_2,$$

then both conditions (ii)a and (ii)b are satisfied.

5.6.1 Numerical Values of N

Under condition (i), we require the smallest integer

N to satisfy

$$\rho^{N-1-m} < \varepsilon_1 .$$

If we let $\varepsilon_1 = 10^{-2}$, we are approximating to the terms π_1 which form 99% of the probability distribution $\underline{\pi}$. For $m = 1$ and $\rho = 0.7$, we obtain $N = 15$; the equivalent value for $\rho = 0.9$ is $N = 46$.

Under condition (ii) we require the smallest N to satisfy

$$c\rho^{N-1}(1-\rho) < \varepsilon_2 .$$

We let $\varepsilon_2 = 10^{-4}$ and $c = 0.05$ (or 5% error in $\theta_N^{(N)}$, the term with greatest proportional error). We require N to satisfy

$$0.05.\rho^{N-1}(1-\rho) < 10^{-4} .$$

For $\rho = 0.7$ we find $N = 16$, and for $\rho = 0.9$, $N = 39$.

A few trial solutions of the system M/M/1 using this method indicated that $N = 20$ for values of ρ not greater than 0.7, and $N = 40$ for $\rho = 0.8$ and 0.9 gave sufficient accuracy in the statistics of interest to be derived from $\underline{\pi}^{(N)}$, and these choices did not involve a prohibitive amount of computing time in the solutions of equations (5.15) and (5.16).

5.7 Queueing Time Distributions

We now derive some properties of the queueing time distributions for the two classes of patient, regular and random, by consid-

eration of the distribution $\underline{\pi}$ of the number in the system just after regular arrival instants.

In many queueing theory analyses, the waiting time distribution is of importance; however in this application it was felt that the queueing time was more important. Most patients do not regard the period of examination or consultation as "wasted"; they wish to derive the maximum benefit from their visit to the clinic, and in general they have no desire to hurry through this stage of the process. In contrast, few patients enjoy the queueing phase before treatment; they can see no effort being expended for their direct benefit, and most people regard any sort of queueing in line as a frustrating waste of time which has to be endured. For these reasons we consider the properties of the queueing time distribution of our model. Waiting times are of interest, but do not lead to such a direct approach to problems where the objectives of patient well-being and departmental efficiency conflict.

5.7.1 Queueing Time for Scheduled Arrivals

We suppose that there are n people in the system (queueing in line and in service) after some given regular arrival of m people together. We consider the queueing time distribution q(t) of the s th. member of this batch.

(i) Mean and Variance of q(t)

The s th. customer of the batch has to queue through the remaining service time of the person currently receiving attention, and also

the complete service of all the people further up the queue. There are (m - s) people in line behind the s th. member, who therefore has to queue during the complete service periods of {n - (m - s) - 2} individuals, and the remaining part x of the service in progress. Using the Markovian property of the service mechanism, we note that x is an exponential variable of mean $b = 1/\sigma$, the service-time mean. Thus the queueing time t of the s th. individual in the batch is the sum of {n - m + s - 1} independent exponential variables of mean b, that is it has an Erlangian distribution with {n - m + s - 1} phases, conditional on n. Summing over n, we obtain the unconditional distribution as

$$q(t)dt = \sum_{n=m}^{\infty} \pi_n \cdot \frac{(1/b)^{n-m+s-1} . t^{n-m+s-2} . e^{-t/b}}{(n-m+s-2)!} . dt \qquad (5.25)$$

Taking expectations, we have

$$E(t) = \int_0^{\infty} t q(t) dt$$

$$= \sum_{n=m}^{\infty} \pi_n \int_0^{\infty} \frac{(1/b)^{n-m+s-1} . t^{n-m+s-1} . e^{-t/b}}{(n-m+s-2)!} . dt$$

$$= \sum_{n=m}^{\infty} \pi_n . b . \frac{(n-m+s-1)!}{(n-m+s-2)!}$$

$$= b \sum_{n=m}^{\infty} \pi_n (n-m+s-1)$$

$$= b \{E(n) - (m-s+1)\} \qquad (5.26)$$

where E(n) is the expectation of n. Similarly we may derive

$$E(t^2) = b\{E(n^2) - E(n)(2m - 2s + 1) + (m - s)(m - s + 1) + s - 1\}$$

and thus

$$Var(t) = b^2\{Var(n) + E(n) - 2(m - s + 1) + m\}$$

(ii) Quantiles of q(t)

Using (5.25) we have that

$$Pr(t > T) = \int_T^\infty q(t)dt$$

$$= \sum_{n=m}^{\infty} \pi_n \cdot \frac{\sigma^{n-m+s-1}}{(n-m+s-2)!} \int_T^\infty t^{n-m+s-2} \cdot e^{-\sigma t} dt$$

$$= \sum_{n=m}^{\infty} \frac{\pi_n}{(n-m+s-2)!} \int_{\sigma T}^\infty u^{n-m+s-2} \cdot e^{-u} du \quad (T \geq 0)$$

We let

$$\Psi(a,x) = \frac{1}{\Gamma(a)} \int_0^\infty e^{-x} \cdot x^{a-1} \cdot dx,$$

the incomplete gamma function. Then

$$Pr(t > T) = \sum_{n=m}^{\infty} \pi_n \{1 - \Psi(n - m + s - 1, \sigma T)\} \qquad (5.27)$$

As the order of the gamma function is integral, we may use the identity

$$\Psi(k,x) = 1 - \{1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \ldots + \frac{x^{k-1}}{(k-1)!}\} e^{-x}$$

$$= 1 - e_{k-1}(x).e^{-x}, \text{ say.} \qquad (5.28)$$

Then we have for $s > 1$,

$$\Pr(t > T) = \sum_{i=m}^{\infty} \pi_i \; e_{i-m+s-2}(\sigma T)e^{-\sigma T} \quad (T \geq 0) \qquad (5.29)$$

For $s = 1$ the result is

$$\Pr(t > T) = \sum_{i=m+1}^{\infty} \pi_i \; e_{i-m+s-2}(\sigma T)e^{-\sigma T} \quad (T > 0)$$

and

$$\Pr(t = 0) = \pi_m \qquad (5.30)$$

The use of identity (5.28), which involves only the partial sums of the exponential function, makes the numerical calculation of the quantiles of $q(t)$ simple and precise; in contrast, the use of tables required with expression (5.27) may be somewhat impractical.

5.7.2. Queueing Time for Unscheduled Arrivals

We consider the state of the system after an unscheduled (or random) arrival at a time u after the last regular bulk arrival. For convenience we take the time origin as the time of the last regular batch, and we let n_t be the number in the system after a time t. By convention, we consider the instant $t = 0$ when there is an arrival at

t, unless otherwise stated. We also let q(v) be the queueing time distribution of the given random arrival. The properties of q(v) and n_t are harder to derive than the corresponding items for regular arrivals; this is because the approximation $\underline{\pi}'$ to $\underline{\pi}$ applies to instants after regular batches.

(i) Mean of q(v)

Once again we use the Markovian property of the service mechanism, and it is clear that

$$E(v) = \int_0^\infty vq(v)dv$$

$$= b\{E(n_{u+}) - 1\}$$

We therefore consider the distribution of n_t.

The expected number in the system just after a regular arrival is

$$E(n_{o+}) = \sum_{i=m}^{\infty} i\,\pi_i$$

We denote the general transition probability from i to j over any time interval (x, x + t), such that 0 < x and x + t < 1, by $p_t(i, j)$; we also let $p_t(j)$ be the probability of j in the system unconditional on i. Then

$$p_t(j) = \sum_{i=m}^{\infty} \pi_i\, p_t(i,j)$$

and with the given random arrival at t = u, we have

$$\Pr(n_{u+} = j + 1) = p_t(j).$$

Then

$$E(n_t) = \sum_{i=m}^{\infty} \pi_i \sum_{j=o}^{\infty} j p_t(i,j)$$

and

$$E(n_{u+}) = E(n_u) + 1$$

Lemma 1.

$E(n_t) \leqslant E(n_{o+})$ for $0 \leqslant t < 1$, with equality only when $t = 0$.

Proof:-

We may write $E(n_{o+})$ as

$$E(n_{o+}) = \sum_{i=m}^{\infty} \pi_i \sum_{j=o}^{\infty} j\, p_o(i,j)$$

$$= \sum_{i=m}^{\infty} \pi_i \sum_{j=o}^{\infty} j.\delta_{ij} ,$$

where δ_{ij} is Kronecker's delta function,

$$= \sum_{i=m}^{\infty} \pi_i.i .$$

From the equilibrium of the system, we have

$$E(n_{1-}) = E(n_{o+}) - m \qquad (5.31)$$

We prove the result by contradiction; suppose there exists some value t^* of t such that $0 \leq t^* < 1$ and $E(n_{t^*}) = E(n_{o+})$,

i.e.
$$\sum_{i=m}^{\infty} \pi_i \sum_{j=o}^{\infty} j\, p_{t^*}(i,j) = \sum_{i=m}^{\infty} i\pi_i \ . \tag{5.32}$$

Comparing coefficients of π_i in (5.32), we must have

$$i = \sum_{j=o}^{\infty} j\, p_{t^*}(i,j) \tag{5.33}$$

for all i such that $m \leq i < \infty$. We also note that the expression on the right-hand side of (5.33) is the expected number in the system at time t^* , given that there were i at time o+ . Also, clearly i = expected number at $t = o+$, or

$$i = \sum_{j=o}^{\infty} j p_o(i,j) \tag{5.34}$$

We equate (5.33) and (5.34) and differentiate with respect to t.

$$\sum_{j=o}^{\infty} j p_t'(i,j) \Big|_{t=t^*} = \sum_{j=o}^{\infty} j p_t'(i,j) \Big|_{t=o+} ,$$

where ' denotes the differentiation with respect to t, and $|_{t=}$ indicates the value of a function at a particular value of t; derivatives from (5.34) are taken from the right. We now take weighted sums over π_i:-

$$\sum_{i=m}^{\infty} \pi_i \sum_{j=o}^{\infty} j p_t'(i,j) \Big|_{t=t^*} = \sum_{i=m}^{\infty} \pi_i \sum_{j=o}^{\infty} j p_t'(i,j) \Big|_{t=o+}$$

or

$$\left.\frac{df(t)}{dt}\right|_{t^*} = \left.\frac{df(t)}{dt}\right|_{o+} = c \qquad (5.35)$$

where c is some constant, and $f(t) = E(n_t)$, a function of t. f(t) is continuous and has derivatives of all orders in the interval (0, 1). If $c > 0$, by continuity there must be a point t** such that $f(t^{**}) = f(o+)$, and which has derivative $f'(t^{**}) \leq 0$. However t** must also satisfy (5.35), giving the contradiction. Similarly $c \not< 0$. Therefore $c = o$. The argument leading to (5.35) applies to any appropriate choice of t*; we must conclude that $E(n_t) = E(n_{o+})$ everywhere in (0,1), which conflicts with the boundary condition (5.31). Hence the result follows.

Lemma 2

$f'(t) \leq 0$ for $0 < t < 1$, where f(t) is as before.

Proof:-

We assume the result is false; we may therefore suppose there exist values t_1 and t_2 of t such that $t_2 > t_1$ and

$$E(n_{t_1}) = E(n_{t_2})$$

and $f'(t) > 0$ for some value of t in the interval (t_1, t_2). Let the distributions of the number in the system at times t_1 and t_2 be $\underline{\pi}^{(1)}$ and $\underline{\pi}^{(2)}$ respectively. Then

$$\sum_{i=o}^{\infty} \pi_i^{(1)}.i = \sum_{i=o}^{\infty} \pi_i^{(2)}.i$$

by assumption. Also

$$\pi_j^{(2)} = \sum_{i=o}^{\infty} \pi_i^{(1)} \, p_{t_2-t_1}(i,j).$$

Therefore

$$\sum_{i=o}^{\infty} \pi_i^{(1)}.i = \sum_{j=o}^{\infty} j \sum_{i=o}^{\infty} \pi_i^{(1)} \cdot p_{t_2-t_1}(i,j) \qquad (5.36)$$

Comparing coefficients of $\pi_i^{(1)}$ in (5.36) gives a contradiction on the boundary condition (5.31) as before.

Combining the results of these two lemmas, we have that

$$E(n_{u+}) \leqslant E(n_{o+}) + 1.$$

The equivalent result for queueing times is

$$E(v) \leqslant bE(n_{o+})$$

= Expected waiting time for regular arrivals.

(ii) Quantiles of q(v)

Given that there are j in the system just after time t = u, the random arrival at n has to queue during the complete service of (j - 2) people, and the remaining service of one person. As before, the distribution of the part service time is exponential of mean b, and the whole queueing time has an E_{j-1} distribution of mean (j - 1)b. Thus the unconditional distribution is

$$q(v)dv = \sum_{i=m}^{\infty} \pi_i \sum_{j=i}^{\infty} p_u(i,j).\frac{(1/b)^{j-1}.v^{j-2}.e^{-v/b}}{(j-2)!}.dv \quad \text{for } v > 0,$$

and

$$q(v) = \sum_{i=m}^{\infty} \pi_i \, p_u(i, 0) \quad (v = 0).$$

For the quantiles,

$$Pr(v > T) = \int_T^{\infty} q(v)dv$$

$$= \sum_{i=m}^{\infty} \pi_i \sum_{j=1}^{\infty} p_u(i,j) \int_T^{\infty} \frac{(1/b)^{j-1}.v^{j-2}.e^{-v/b}}{(j-2)!}.dv$$

$$= \sum_{i=m}^{\infty} \pi_i \sum_{j=1}^{\infty} p_u(i,j) \{1 - \Psi(j, T/b)\}$$

$$= \sum_{i=m}^{\infty} \pi_i \sum_{j=1}^{\infty} p_u(i,j) e_{j-1}(T/b)e^{-T/b} \qquad (T > 0)$$

and

$$pr(v = 0) = \sum_{i=m}^{\infty} \pi_i \; p_u(i,0) \tag{5.37}$$

The above distribution is conditional on the random arrival considered being at time t = u; to obtain the general distributions for all random arrivals, equations (5.37) must be integrated over u ranging from 0 to 1. This is rather difficult numerically, but by considering again the underlying mechanism of the

system, certain statistics may be derived.

The random arrival time u is by definition distributed over $[0,1)$ in any given epoch between bulk regular arrivals. Thus the unconditional expectation of v will be the same as that for a simple queue of intensity ρ , i.e.

$$E_u(v) = b/(1 - \rho) ,$$

where E_u indicates an expectation of v arising from (5.37) integrated over u. Also implied from the arguments about n_t, we have that $E(v|u)$ is a monotonic decreasing function of u.

We now consider the variance of v. Firstly from the equilibrium of the system we note that the distribution of n_{0+} is merely a translation by m of the distribution of n_{1-}. Thus we have

$$\mathrm{Var}(v|u = 0+) = \mathrm{Var}(v|u = 1-).$$

Afetr a bulk arrival, $E(n_t)$ decreases monotonically; if there were no further bulk arrivals this function would tend to a limit equal to the expected number after an arrival in the simple queue of intensity ρ_s , i.e. $1/(1 - \rho_s)$.

Now we consider two initial values i_1 and i_2 of n_{t_1} , and we investigate the behaviour of $n_{\Delta + t_1}$, where Δ is a finite time. Suppose i_1 is very large; then n_t will usually decrease over $(t_1, t_1 + \Delta)$, and rarely will it equal zero if Δ is not too large. If i_2 is taken as small, any sample behaviour of n_t will be subject to the same increases as before (arising from random arrivals which are independent of n_{t_1}); n_t can never be less than zero, and $n_{t_1 + \Delta}$ has a non-zero

probability of being zero. Thus in general terms a larger value of n_t will yield a larger value of $\mathrm{Var}(n_{t+\Delta})$ for an appropriate value of Δ. Combining all these arguments, we may deduce the behaviour of Var(v) over the time interval $[0,1)$, and the form of its graph is shown in Figure 5.6. The expression

$$\frac{\rho_s b^2}{(1-\rho_s)^2}$$

is the limit variance, equal to the variance for the simple queue of intensity ρ_s.

5.8. Numerical Results

5.8.1. Expectation of n_{o+}

In the previous notation, n_{o+} corresponds to π'_n; the vector

$$\underline{\pi}' = (\pi'_m, \ldots . \pi'_{m+N-1})$$

is the solution of the set of equations (5.24), which approximates to $\underline{\pi}$, the true distribution of the number n in the system just after a regular bulk arrival. Table 5.7 gives the values of E(n) - m as estimated by this method, over the ranges of interest of the system parameters. The table may be used in (5.26) to estimate the expected queueing time of regular arrivals.

For small values of r and ρ, and large m, the system behaves very similarly to a simple queue of intensity ρ_s. Thus in the appropriate limits for these

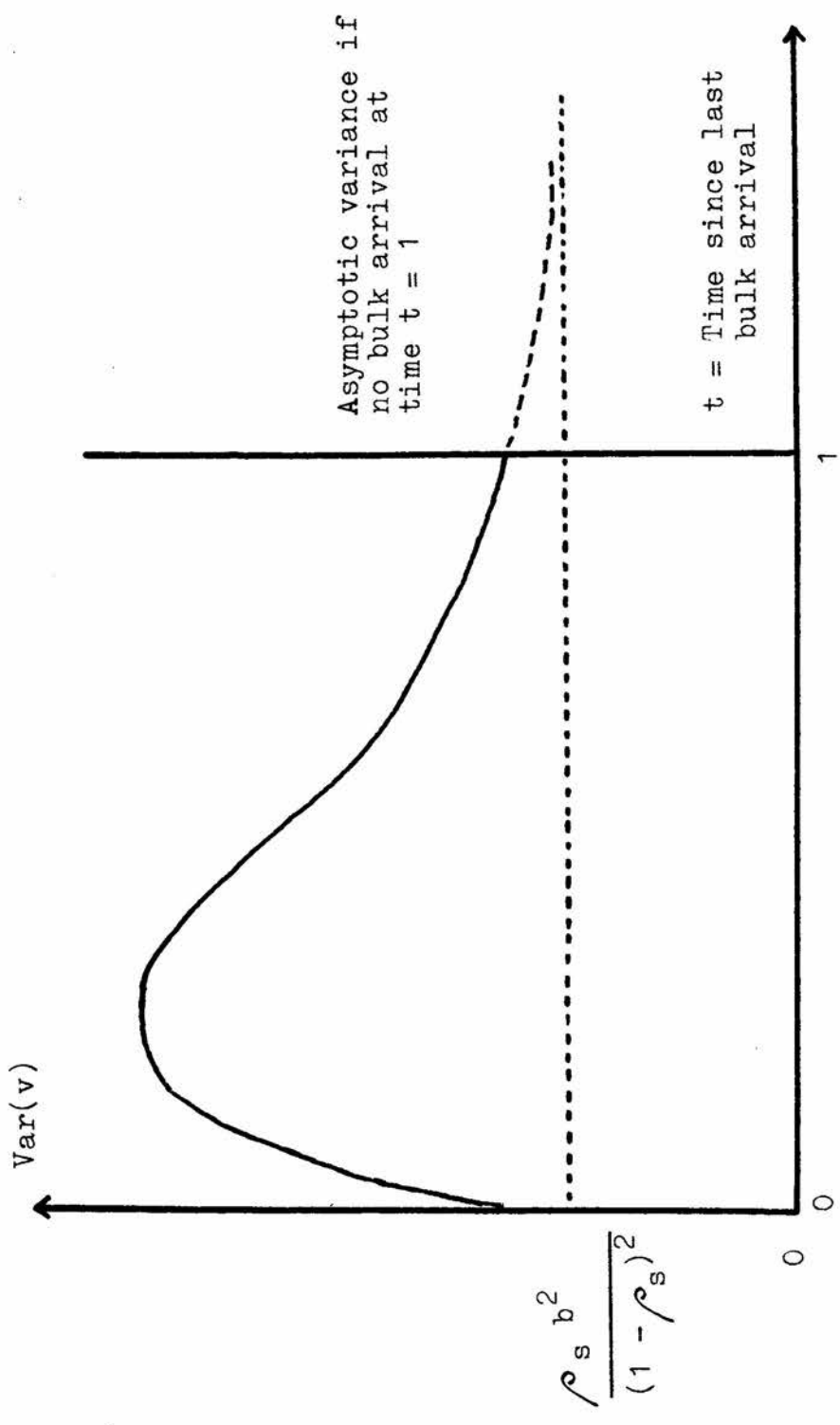


FIGURE 5.6 VARIANCE OF QUEUEING TIME FOR RANDOM ARRIVALS (Sketch)

TABLE 5.7 E(n) - m

	r	0.2	0.5	1.0	1.5	2.0	5.0
ρ	m						
	1	0.509	0.404	0.323	0.279	0.251	0.184
	2	0.502	0.376	0.278	0.228	0.196	0.123
0.4	3	0.501	0.368	0.261	0.207	0.173	0.097
	4	0.501	0.366	0.256	0.198	0.163	0.084
	5	0.501	0.365	0.253	0.195	0.159	0.079
	1	0.762	0.632	0.528	0.471	0.433	0.343
	2	0.729	0.565	0.433	0.380	0.341	0.247
0.5	3	0.719	0.534	0.399	0.331	0.290	0.193
	4	0.717	0.519	0.373	0.302	0.259	0.160
	5	0.716	0.511	0.359	0.284	0.240	0.138
	1	1.164	1.002	0.865	0.787	0.734	0.606
	2	1.084	0.887	0.738	0.656	0.604	0.475
0.6	3	1.046	0.818	0.657	0.573	0.519	0.392
	4	1.025	0.773	0.599	0.513	0.459	0.332
	5	1.015	0.742	0.560	0.470	0.414	0.288
	1	1.868	1.654	1.457	1.340	1.263	1.069
	2	1.729	1.489	1.287	1.172	1.097	0.909
0.7	3	1.639	1.372	1.166	1.052	0.978	0.794
	4	1.576	1.284	1.072	0.958	0.885	0.706
	5	1.532	1.213	0.996	0.882	0.810	0.636

TABLE 5.7 (continued)

ρ	r / m	0.2	0.5	1.0	1.5	2.0	5.0
0.8	1	3.358	3.018	2.693	2.498	2.360	2.021
	2	3.145	2.798	2.483	2.294	2.158	1.842
	3	2.986	2.640	2.316	2.129	2.006	1.700
	4	2.857	2.495	2.181	2.001	1.882	1.580
	5	2.752	2.374	2.067	1.890	1.774	1.481
0.9	1	7.545	6.934	6.298	5.881	5.612	4.904
	2	7.295	6.684	6.044	5.671	5.385	4.700
	3	7.087	6.476	5.864	5.740	5.213	4.537
	4	6.895	6.291	5.683	5.317	5.059	4.393
	5	6.895	6.291	5.683	5.317	5.059	4.393

parameters we would expect the distribution $\underline{\pi}$ to become geometric on the integers m, m + 1, ... , with a mean of $m + \rho_s/(1 - \rho_s)$. Table 5.8 gives the values of $\rho_s/(1 - \rho_s)$ as determined by ρ and r, using the results from Table 5.3.

5.8.2. Quantiles of q(v) for regular arrivals

Using identities (5.29) and (5.30) quantiles of q(v) may be evaluated for each member of a batch of m regular arrivals. Figures 5.7, 5.8 and 5.9 show some typical distribution functions which result, interpolated from the estimated probabilities $Pr(v \leq cb)$, where b is the mean service time, and c takes the values 1, 2, 3, 4 and 5. For s = 1, the distribution function starts at a non-zero value at v = 0, reflecting the fact that the first member of a batch may not have to queue at all.

In practice the members of a batch are served in random order. Even in clinics where patients with the same appointment time are ordered in a list, there is often no guarantee that the order of service will correspond to the list order also the original composition of a list is subject to a large number of variable influences. We may therefore be interested in the probability that v exceeds certain values, averaged over all members of a batch. Tables 5.9 and 5.10 give the probabilities that $v > cb$ with c = 2 and 5 respectively.

5.8.3. Comparison with r = o and r = ∞

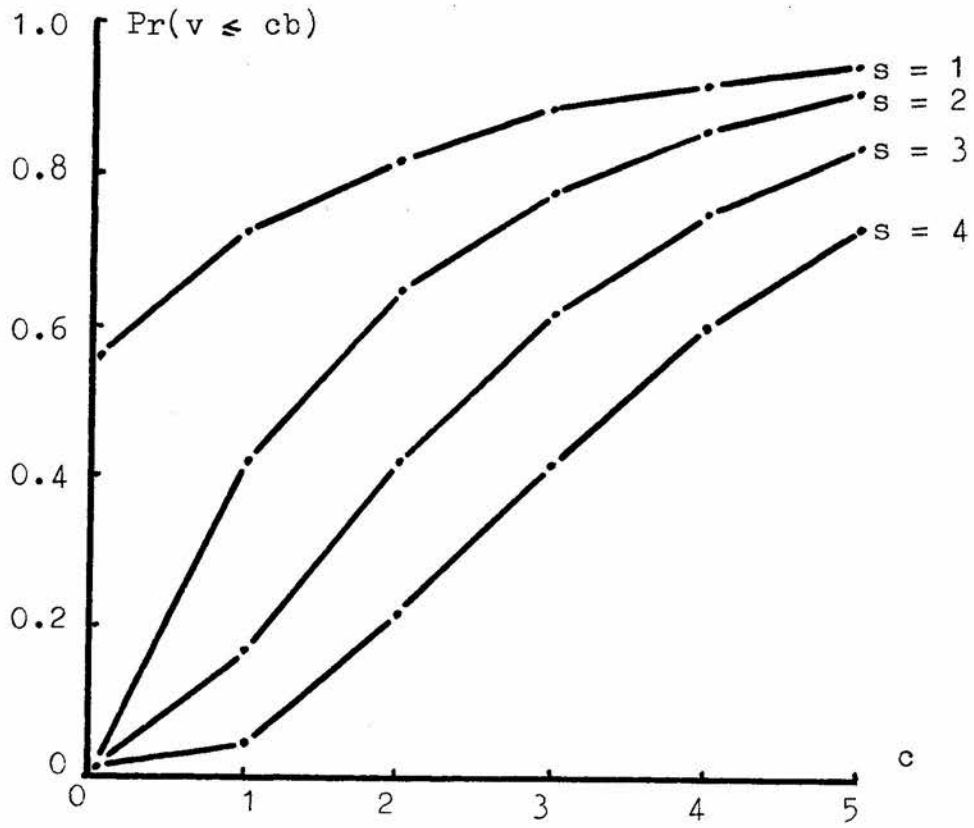
(i) r = o

r = o gives the system M/M/1 with the same intensity. The queueing-time distribution is known as

TABLE 5.8 VALUES OF $\rho_s/(1-\rho_s)$

ρ \ r	0.2	0.5	1.0	1.5	2.0	5.0
0.4	0.50	0.36	0.25	0.19	0.15	0.07
0.5	0.72	0.50	0.33	0.25	0.20	0.09
0.6	1.00	0.67	0.43	0.32	0.25	0.11
0.7	1.39	0.88	0.54	0.39	0.30	0.13
0.8	2.00	1.14	0.67	0.47	0.36	0.15
0.9	3.00	1.50	0.82	0.56	0.43	0.18

FIGURE 5.7 CUMULATIVE DISTRIBUTION FUNCTION OF QUEUEING TIME FOR VARIOUS MEMBERS OF A BATCH ARRIVAL: $r = 1.0$, $m = 4$, $\rho = 0.7$



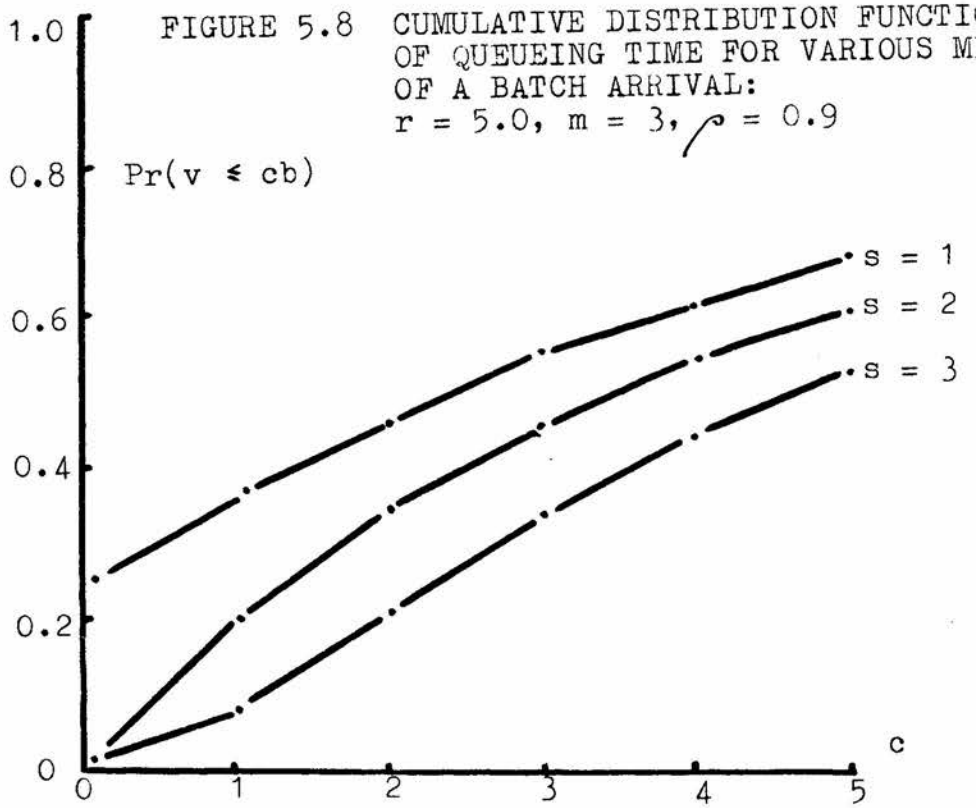


FIGURE 5.8 CUMULATIVE DISTRIBUTION FUNCTION OF QUEUEING TIME FOR VARIOUS MEMBERS OF A BATCH ARRIVAL:
r = 5.0, m = 3, ρ = 0.9

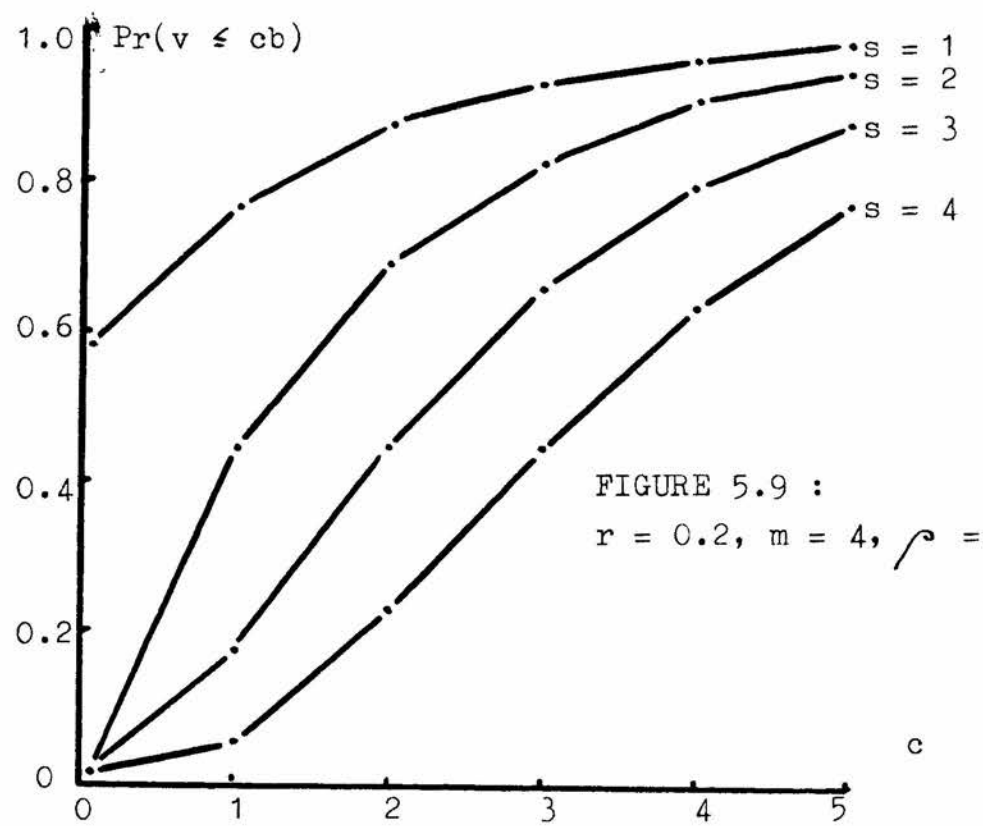


FIGURE 5.9 :
r = 0.2, m = 4, ρ = 0.5

TABLE 5.9 Pr(v > 2b)

	r	0.2	0.5	1.0	1.5	2.0	5.0
ρ	m						
	1	0.089	0.070	0.055	0.047	0.042	0.030
	2	0.176	0.148	0.127	0.116	0.109	0.094
0.4	3	0.291	0.262	0.239	0.227	0.219	0.202
	4	0.405	0.380	0.358	0.346	0.339	0.322
	5	0.502	0.480	0.461	0.515	0.444	0.430
	1	0.138	0.114	0.095	0.084	0.077	0.059
	2	0.223	0.188	0.162	0.149	0.141	0.121
0.5	3	0.334	0.296	0.267	0.253	0.244	0.223
	4	0.443	0.407	0.380	0.366	0.357	0.337
	5	0.534	0.503	0.478	0.466	0.458	0.440
	1	0.210	0.182	0.158	0.144	0.135	0.110
	2	0.288	0.249	0.220	0.205	0.195	0.170
0.6	3	0.389	0.346	0.315	0.299	0.290	0.264
	4	0.489	0.447	0.417	0.401	0.391	0.368
	5	0.573	0.535	0.507	0.492	0.484	0.463
	1	0.316	0.286	0.258	0.240	0.228	0.195
	2	0.381	0.343	0.312	0.295	0.283	0.253
0.7	3	0.468	0.425	0.393	0.376	0.365	0.336
	4	0.553	0.510	0.479	0.463	0.453	0.428
	5	0.625	0.585	0.556	0.542	0.532	0.510

TABLE 5.9 (continued)

ρ	r / m	0.2	0.5	1.0	1.5	2.0	5.0
	1	0.470	0.441	0.411	0.392	0.379	0.341
	2	0.519	0.484	0.455	0.437	0.425	0.392
0.8	3	0.583	0.546	0.517	0.500	0.489	0.460
	4	0.647	0.610	0.583	0.567	0.557	0.532
	5	0.761	0.726	0.641	0.626	0.619	0.596
	1	0.687	0.666	0.644	0.629	0.618	0.586
	2	0.715	0.692	0.670	0.658	0.648	0.620
0.9	3	0.751	0.728	0.708	0.695	0.687	0.663
	4	0.787	0.764	0.746	0.735	0.727	0.707
	5	0.818	0.796	0.779	0.770	0.763	0.643

TABLE 5.10 Pr(v > 5b)

ρ	r / m	0.2	0.5	1.0	1.5	2.0	5.0
	1	0.013	0.009	0.007	0.005	0.005	0.003
	2	0.024	0.017	0.013	0.011	0.010	0.007
0.4	3	0.047	0.037	0.030	0.027	0.025	0.021
	4	0.085	0.072	0.063	0.058	0.055	0.049
	5	0.138	0.123	0.112	0.106	0.102	0.095
	1	0.026	0.020	0.016	0.013	0.011	0.007
	20	0.040	0.030	0.023	0.020	0.018	0.013
0.5	3	0.067	0.051	0.041	0.036	0.033	0.027
	4	0.108	0.087	0.074	0.068	0.064	0.056
	5	0.163	0.139	0.123	0.115	0.111	0.101
	1	0.056	0.046	0.036	0.031	0.027	0.019
	2	0.070	0.056	0.046	0.040	0.036	0.027
0.6	3	0.099	0.078	0.065	0.058	0.053	0.043
	4	0.142	0.114	0.098	0.090	0.084	0.073
	5	0.197	0.166	0.146	0.136	0.130	0.117
	1	0.117	0.100	0.083	0.073	0.066	0.049
	2	0.132	0.113	0.095	0.085	0.078	0.061
0.7	3	0.160	0.135	0.116	0.106	0.099	0.081
	4	0.201	0.170	0.149	0.138	0.131	0.113
	5	0.254	0.198	0.195	0.183	0.176	0.157

TABLE 5.10 (continued)

ρ	r / m	0.2	0.5	1.0	1.5	2.0	5.0
0.8	1	0.243	0.218	0.191	0.174	0.162	0.131
	2	0.257	0.232	0.206	0.190	0.178	0.148
	3	0.282	0.254	0.228	0.212	0.201	0.172
	4	0.317	0.285	0.260	0.244	0.234	0.206
	5	0.362	0.326	0.300	0.285	0.275	0.249
0.9	1	0.492	0.465	0.434	0.413	0.398	0.356
	2	0.503	0.478	0.449	0.430	0.415	0.374
	3	0.521	0.495	0.469	0.450	0.446	0.400
	4	0.544	0.518	0.493	0.476	0.464	0.430
	5	0.573	0.546	0.522	0.507	0.496	0.464

$$pr(v = o) = 1 - \rho$$

$$q(v)dv = \rho(\sigma - \alpha)e^{-(\sigma - \alpha)v}dv \quad ,$$

where α and σ are the arrival and service rates.

We may derive that

$$pr(v > c/\sigma) = pr(v > cb)$$

$$= \rho \exp\left[- c(1 - \rho)\right] .$$

Table 5.11 gives these probabilities for c = 2 and 5 as before.

(ii) <u>r = ∞</u>

When r = ∞ we have the system $D_m/M/1$ with the same intensity ρ. For m = 1 we have from (5.5) that the probability that an arriving unit finds n units already in the system is

$$q_n = (1 - e^{-y_o})\, e^{-ny_o} \quad ,$$

where y_o is the positive real cost of

$$\rho = (1 - e^{-y})/y \ .$$

The queueing time distribution is then

$$q(v)dv = \sum_{n=1}^{\infty} q_n \frac{\sigma^n . e^{-\sigma t} . v^{n-1}}{(n-1)!} . dv$$

$$= \sigma e^{-y_o}(1 - e^{-y_o}) \exp\{- \sigma(1 - e^{-y_o})v\}dv$$

and

$$pr(v = o) = 1 - e^{-y_o}$$

Thus

$$Pr(v > cb) = \int_{cb}^{\infty} \alpha e^{-\beta v} dv ,$$

where α and β are constants

$$= \exp\{-y_o (1 + c\rho)\} \qquad (5.38)$$

y_o was evaluated for appropriate values of ρ by a Newton-Raphson process, and the results are given in Table 5.12. Using these figures, pr(v > c b) was calculated for c = 2 and 5 as before, and the results are shown in Table 5.13.

5.9. Approach to Equilibrium of the Model

Throughout the above argument we have made the assumption that equilibrium conditions exist, and all the results derived are based on that premise. In reality, the patient input to a clinic will probably deviate from a constant one because of changes in the arrival rates from different sources during a session, tea breaks, and other local detail which yield a system probably impossible to model exactly by analysis. The assumption of equilibrium is one which leads to a prediction of the usual overall behaviour of the system, without too much analytic complexity.

Even in the unlikely event that the input is constant in the manner described,

TABLE 5.11 $Pr(v > cb)$ FOR $r = 0$

ρ \ c	2	5
0.4	0.121	0.020
0.5	0.184	0.041
0.6	0.270	0.081
0.7	0.384	0.156
0.8	0.536	0.294
0.9	0.737	0.546

TABLE 5.12 VALUES OF y_o

ρ	0.4	0.5	0.6	0.7	0.8	0.9
y_o	2.23	1.59	1.13	0.76	0.46	0.21

TABLE 5.13 $Pr(v > cb)$ FOR $r = \infty$

ρ \ c	2	5
0.4	0.018	0.001
0.5	0.041	0.004
0.6	0.084	0.011
0.7	0.161	0.033
0.8	0.299	0.098
0.9	0.548	0.307

the system may still exhibit time-dependent behaviour as the session is of finite duration. Strictly speaking, equilibrium conditions will only arise if either the initial parameter values are carefully chosen, or if the system has been running for an infinitely long time and has settled to a behaviour independent of the starting values. We must therefore investigate how rapidly our model system approaches equilibrium for initial values likely to occur in practice.

In an attempt to equalise the expected queueing times of patients at the beginning and end of a session, many clinics have adopted, with some theoretical justification (cf. Bailey 1955), a policy of calling rather more patients by appointment at the start of a clinic than in a normal batch. We will assume in this investigation that we have regular batch arrivals at unit time intervals at times $t = 1, 2, 3, \ldots$, and the initial distribution of the number in the system will be specified later. We now let the distribution of the number in the system just after the n th. batch arrival (i.e. at time $t = n + o$) be

$$\underline{\pi}^{(n)} = (\pi_m^{(n)}, \pi_{m+1}^{(n)}, \ldots) \qquad (n \geq 1)$$

We have as before that if there are i in the system just after the n th. batch, then

$$\pi_{j+m}^{(n+1)} = p(i,j) \qquad j = 0, 1, 2, \ldots$$

Unconditionally

$$\pi_{j+m}^{(n+1)} = \sum_{i=m}^{\infty} \pi_i^{(n)} . p(i,j)$$

and in matrix form we have a system of equations

$$\underline{\pi}^{(n+1)^T} = \underline{\pi}^{(n)^T} \cdot \underline{P} \tag{5.39}$$

where

$$\underline{\pi}^{(n)^T} = (\pi_m^{(n)}, \pi_{m+1}^{(n)}, \ldots.)$$

and $\underline{P}$ is the infinite matrix of transition probabilities as before.

Using (5.38) repeatedly we have that

$$\begin{aligned} \underline{\pi}^{(n)} &= \underline{P}^T \cdot \underline{\pi}^{(n-1)} \\ &= (\underline{P}^T)^{n-1} \cdot \underline{\pi}^{(1)} \end{aligned} \tag{5.40}$$

Let

$$\underline{\pi}^{(0)^T} = (\pi_0^{(0)}, \pi_1^{(0)}, \pi_2^{(0)}, \ldots.)$$

be the initial distribution at $t = 0$, and let $\underline{Q}$ be the transition probability matrix

$$\begin{pmatrix} p(0,0) & ; & p(0,1) & ; & p(0,2) & ; & \ldots. \\ p(1,0) & ; & p(1,1) & ; & p(1,2) & ; & \ldots. \\ \vdots & & \vdots & & \vdots & & \end{pmatrix}$$

Then

$$\underline{\pi}^{(1)} = \underline{Q}^T \cdot \underline{\pi}^{(0)}$$

and we have that

$$\underline{\pi}^{(n)} = (\underline{P}^T)^{n-1} \cdot \underline{Q}^T \cdot \underline{\pi}^{(0)} \tag{5.41}$$

If a equilibrium distribution $\underline{\pi}$ exists, then $\underline{\pi}^{(n)}$ tends, in some sense, to $\underline{\pi}$ as n tends to infinity. As $\underline{\pi}$ is independent of $\underline{\pi}^{(0)}$, we must have from (5.41) that the rows of $(\underline{P}^T)^{n-1}$ become almost equal as n becomes large. Formally, let e_{ij}^n be the (i,j) th. element of $(\underline{P}^T)^{n-1}$. Then we have that

$$e_{ij}^n \to e_i \quad \text{as} \quad n \to \infty$$

independently of j. We also let f_{ij}^n be the (i,j) th. element of $(\underline{P}^T)^{n-1}.\underline{Q}^T$. Then

$$f_{ij}^n = \sum_{j=1}^{\infty} e_{ij}^n \cdot p(i-1, j-1).$$

As

$$\sum_{n=0}^{\infty} p(m,n) = 1 \quad \text{for any } m \geq 0, \text{ and}$$

$$\sum_{i=0}^{\infty} \pi_i^{(0)} = 1,$$

we must have that

$$f_{ij}^n \to f_i \quad \text{as} \quad n \to \infty,$$

where f_i is some constant limit, equal to π_{i+m}.

5.9.1 Numerical Computation

It is to be expected that a "reasonable" starting distribution $\underline{\pi}^{(0)}$ will not yield too extreme behaviour before approaching equilibrium. For numerical work it is necessary to use finite approximations to $\underline{\pi}^{(n)}$, $\underline{P}$ and $\underline{Q}$

as before. For "reasonable" distributions $\underline{\pi}^{(0)}$ it was decided to use the same number of terms N as previously with the same parameter values; thus we assume that $\underline{\pi}^{(n)}$ is adequately described by a distribution

$$\underline{\pi}^{(n)'} = (\pi_m^{(n)'}, \pi_{m+1}^{(n)'}, \ldots \pi_{m+N-1}^{(n)'})$$

which satisfies similar criteria to those of the approximation $\underline{\pi}'$ to $\underline{\pi}$. Implicit here is the assumption that truncation in this way does not lead to increasing errors of significant size in the successive derivations of $\underline{\pi}^{(n)}$ from (5.40). Similar to the solution of the equilibrium equations, we define truncated matrices $\underline{P}'$ and $\underline{Q}'$ as approximations to $\underline{P}$ and $\underline{Q}$ for numerical work, defined as

$$\underline{P}' = \begin{bmatrix} p(m,0) & ; p(m,1) & ; \ldots. & p(m,N-1) \\ p(m+1,0) & ; p(m+1,1) & ; \ldots. & p(m+1,N-1) \\ \vdots & \vdots & & \vdots \\ p(m+N-1,0) & ; p(m+N-1,1) & ; \ldots. & p(m+N-1,N-1) \end{bmatrix}$$

and $\underline{Q}' =$

$$\begin{bmatrix} p(0,0) & ; & p(0,1) & ; \ldots. & p(0,N-1) \\ p(1,0) & ; & p(1,1) & ; \ldots. & p(1,N-1) \\ \vdots & & \vdots & & \vdots \\ p(m-1,0) & ; & p(m-1,1) & ; \ldots. & p(m-1,N-1) \\ & & \underline{P}' & & \end{bmatrix}$$

For consistency, the dimension of $\underline{\pi}^{(0)'}$ is taken as (m + N). This approximation will be satisfactory for initial distributions $\underline{\pi}^{(0)}$ which do not yield any terms

$\pi_j^{(n)}$, $j > m + N$, of a size significantly different from zero. However if we had $\pi_i^{(0)} = \delta_{i,\, m+N-1}$, then $\pi_{m+N}^{(1)}$ would not be negligible. We may adapt the method by adding columns to $\underline{Q}'$ and obtaining a better approximation to $\underline{\pi}^{(1)}$; for the finite matrix multiplication, the dimensions of $\underline{P}'$ must also be increased.

Using the above method, it was possible to derive the distributions $\underline{\pi}^{(n)'}$ successively from a given $\underline{\pi}^{(0)}$ from (5.40). If our approximations $\underline{P}'$ and $\underline{Q}'$ are adequate, $\underline{\pi}^{(n)'}$ will tend in some sense to $\underline{\pi}'$, which is in turn an approximation to $\underline{\pi}$ of known accuracy. It was decided to evaluate the principle statistic of interest the mean, from each distribution $\underline{\pi}^{(n)'}$ until this mean had converged sufficiently close to the mean of $\underline{\pi}'$, which had been evaluated previously from the equilibrium equations. When this was done, it was found that the variance of the distributions tended to the variance of $\underline{\pi}'$ at about the same rate as the mean tended towards its limit. Ideally when comparing two distributions in this way, we would prefer a knowledge of all the moments of both. However in this investigation we do not require an exact determination of the rate of convergence, but rather comparisons of the rates for different sets of parameter values, together with a rough indication of the order of magnitude of the convergence rates. Therefore in practice it was decided to use the convergence of the mean as the indication of the closeness of the system to the equilibrium state.

A few sets of the parameter triple (m,r,ρ) were selected, and the system solved using the numerical method. Four starting distributions $\underline{\pi}^{(0)}$ were used for each; and these were

(i) $\pi_i^{(0)} = \delta_{io}$ $\quad i = 0, 1, 2, \ldots$

(ii) $\pi_i^{(0)} = \delta_{i5}$ $\quad i = 0,1,2,\ldots$

(iii) $\pi_i^{(0)} = \frac{1}{2}(\delta_{io} + \delta_{i5})$ $\quad i = 0,1,2,\ldots$

(iv) $\pi_i^{(0)} = 1/6\ (\delta_{i0} + \delta_{i1} + \ldots + \delta_{i5})$ $i = 0,1,2,\ldots$

Distribution (i) represents a clinic with no appointment patient at the start, and (ii) a clinic with five patients present at the start. (iii) is a mixture of the two, and (iv) is a variation on this, both with a variable initial number.

In situations where there is more than one patient present at the start of the session, we may regard the model as being almost equivalent to a clinic where the doctor arrives late, thus causing a build-up of the early patients. As an example, suppose we have single patients arriving by appointments at five minute intervals, and that there are three patients called at the start; then this is equivalent to a system where there are single patients at each appointment time, one patient present at the start, and the doctor arriving ten minutes late. When there are random arrivals in addition, then the two systems are not exactly equivalent, but the approximation is good for high values of r (i.e. when there is a majority of appointment arrivals). As we shall see more fully in Chapter 6, the initial number present has a crucial influence on the subsequent behaviour of the queueing system.

Table 5.14 shows the number of distributions $\underline{\pi}^{(n)'}$ for which the mean was evaluated before this statistic had converged to within 20%, 10% and 5% of the mean of $\underline{\pi}'$. In other words the table shows the number of regular batch arrivals after which we consider the system to be in equilibrium. We note that there is more rapid convergence with large values of m, and small values of r and ρ. In most of the cases shown, equilibrium is reached with the 20% accuracy condition is not more than six time units. With most examinations taking five minutes or less on average, this means that most clinics with a constant input as described would

TABLE 5.14 NUMBER OF BATCH ARRIVALS BEFORE CONVERGENCE TO EQUILIBRIUM OF $\underline{\pi}^{(n)}$

Initial distribution $\underline{\pi}^{(0)}$			(i)			(ii)			(iii)			(iv)		
Percentage difference of mean of $\underline{\pi}^{(n)}$ from mean of $\underline{\pi}$			20	10	5	20	10	5	20	10	5	20	10	5
m	r	ρ												
1	0.2	0.6	1	2	2	2	2	3	0	1	2	2	3	4
1	1.0	0.4	1	2	2	2	2	4	2	3	3	3	4	5
1	1.0	0.6	2	3	5	5	7	9	3	4	6	3	5	7
1	1.0	0.8	3	4	7	6	8	10	3	5	7	4	6	9
1	2.0	0.6	2	4	6	5	8	9	4	6	7	3	5	8
1	5.0	0.4	1	1	2	4	5	6	4	4	5	3	3	4
1	5.0	0.6	2	4	6	5	8	10	4	7	8	6	8	11
3	0.2	0.6	1	1	1	1	1	1	1	1	1	0	1	1
3	1.0	0.6	1	1	2	1	1	2	1	1	1	1	2	2
3	2.0	0.7	1	2	4	1	2	2	1	1	2	0	2	3
5	1.0	0.6	1	1	1	0	1	1	1	1	1	0	1	1
5	5.0	0.6	1	1	1	0	0	1	1	1	1	0	1	1

reach equilibrium (by this criterion) after 30 minutes. Most sessions last in practice about 2½ to 3 hours, and so we may confidently assume equilibrium for most of the session under such a regime. We must be a little more cautious in situations where ρ and r are high, as these systems converge much more slowly.

5.10 Model (VIII): $(M + D_m)/E_k/1$

This system may be dealt with in principle in a manner similar to that for $(M + D_m)/M/1$; we now simply construct equations for the number of "phases" in the system at any time. A customer's service time may be regarded as the sum of k independent exponential phases, and so at each regular arrival time there is an input of mk phases. We denote by $p^*(i,j)$ the phase transition probability. The equilibrium equations equivalent to (5.12) for the previous model are

$$\pi^*_{j + ms} = \sum_{i = ms}^{\infty} \pi_i \, p^*(i,j) \quad j = 0,1,2,.... \tag{5.42}$$

where

$$\underline{\pi}^{*T} = (\pi^*_{ms}, \pi^*_{ms + 1}, \ldots\ldots)$$

is the distribution of the number of phases in the system just after regular arrivals. By convention we regard phases whose service is completed as having passed out of the system; this is simply for mathematical convenience, regardless of the fact that such phases are associated with a patient whose later phases still await service completion. Similarly to before, we define $\underline{P}^*$ as

$$\underline{P}^* = \begin{bmatrix} p^*(ms,0) & ; & p^*(ms,1) & ; & \cdots \\ p^*(ms+1,0) & ; & p^*(ms+1,1) & ; & \cdots \\ \vdots & & & & \end{bmatrix}$$

The system of equations (5.42) may be written as

$$\underline{\pi}^{*T} . (\underline{I} - \underline{P}^*) = \underline{0} .$$

We also have the normalising condition

$$\sum_{i=ms}^{\infty} \pi_i^* = 1 .$$

Thus we have essentially the same system of equations as (5.13) and (5.14) previously. To obtain the distribution $\underline{\pi}$ of the number of people in the system, we remark that "n phases" corresponds to "$\lceil n/k \rceil + 1$ people" in the system, where $\lceil i \rceil$ is the largest integer less than i. Thus

$$\pi_i = \pi^*_{(i-1)m+1} + \pi^*_{(i-1)m+2} + \cdots + \pi^*_{im} \quad i = 1,2,\ldots$$

and

$$\pi_0 = \pi_0^*$$

Queueing and waiting time distributions may be deduced from $\underline{\pi}^*$ by the previous method.

6. Simulation Models of a Clinic Queueing System

6.1 Introduction

In the previous chapter a numerical method is developed for the distribution of the number of units in the queueing system $(M + D_m)/M/1$; the method is extendable in principle to describe the system $(M + D_m)/E_k/1$, a more general model of a clinic. The treatment is mainly for an equilibrium state, but time-dependent solutions are derived, together with an indication of the rate of convergence to equilibrium for certain starting values. It was possible to derive percentiles of the queueing time distributions for both appointment and non-appointment patients.

Although useful as a descriptive tool which yields quantifications of some of the statistics of interest, theoretical models of this kind do have a number of failings and limitations. Firstly, we are limited mathematically to queueing systems which have a traffic intensity less than one. Secondly, we have assumed that our queueing systems are in an equilibrium or steady-state condition, and that this is a valid description of the real clinic. Thirdly, the theoretical models require the numerical solution of the distribution of the number in the system, given in sections 5.4 to 5.8, and this method must be used separately on a different set of equations for each set of parameter values considered. Lastly, in the theoretical model we adopted the same distribution of service times for all patients.

In practice the traffic intensity in a clinic is often close to unity; although the rates of patient arrivals and service completions may vary considerably from time to time, on average patients arrive at roughly the

rate at which the clinic can treat them. A clinic of two or three hours duration is a finite realisation of a queueing system and even by a careful choice of the initial number in the system it may not be possible to induce an equilibrium behaviour. In particular when the traffic intensity is not less than one, there can be no steady-state; in this situation the best we can do is to equalise, say, the expected waiting times of patients at the beginning and end of the session. As we saw in section 4.6.1 there are substantial differences between the service time distributions for patients of different ages, mobility and origin; simulations taking account of this are described later in this chapter.

By using a simulation model we may remove some of the assumptions made in the theoretical study, and obtain a more refined description of the real department. Further empirical details may be included in a model which would probably be impossible to analyse theoretically.

6.2 Description of Simulation Model

A system belonging to the family of queueing models $(M + D_m)/E_k/1$ requires four parameters of specification; these are the traffic intensity ρ, the ratio r of regular and random arrivals, the size m of the batches of regular arrivals, and the parameter k of the service time distribution. For the simulations in the first part of this chapter we will assume a simulation model of the same basic type, with the same Erlangian service time distribution for all patients. We will use an arrival mechanism with a stream of single random patients mixed with a stream of appointment patients in blocks at regular time intervals. The queue discipline will again be "first in, first out". As we are now considering a finite realisation of the queueing system (as opposed to the "infinite" steady-state realisation

of Chapter 5), we must specify two extra parameters, the total number N of patients served in the whole clinic session (referred to as the clinic size), and the number I_o of patients present at the beginning of the session.

In the previous chapter the statistics concerning the waiting and queueing times of patients were in terms of the mean service time, which was taken for convenience as the time unit. Here we will denote the mean service time by b; for the presentation of the numerical results of simulations, a value of b is assumed, but again a scale transformation will give results applicable to systems with different mean service times.

The appointment arrivals are first in one batch of I_o at time 0, and then in batches of m at times c, 2c, 3c, ..., where c is the appointment interval. The parameters r and ρ apply to the periods following the first batch arrival; r is thus the expected ratio of the rates of arrivals of regular and random patients in any time interval not including the origin, and similarly ρ is the expected total traffic intensity for the corresponding period.

6.3 Computer Program for Simulations

Figure 6.1 shows a simplified version of the logic used in the computer simulation program; for clarity, many additional "labelling" variables have been omitted. The program is written in terms of discrete time intervals, and the control moves around the program according to a sequence of events, which are principally either arrivals or departures of patients. At certain times deviations are made from the "event" steps of the program, in order to extract statistics of interest from the simulated realisation of the queueing process.

FIGURE 6.1 REPRESENTATION OF LOGIC OF COMPUTER SIMULATION PROGRAM.

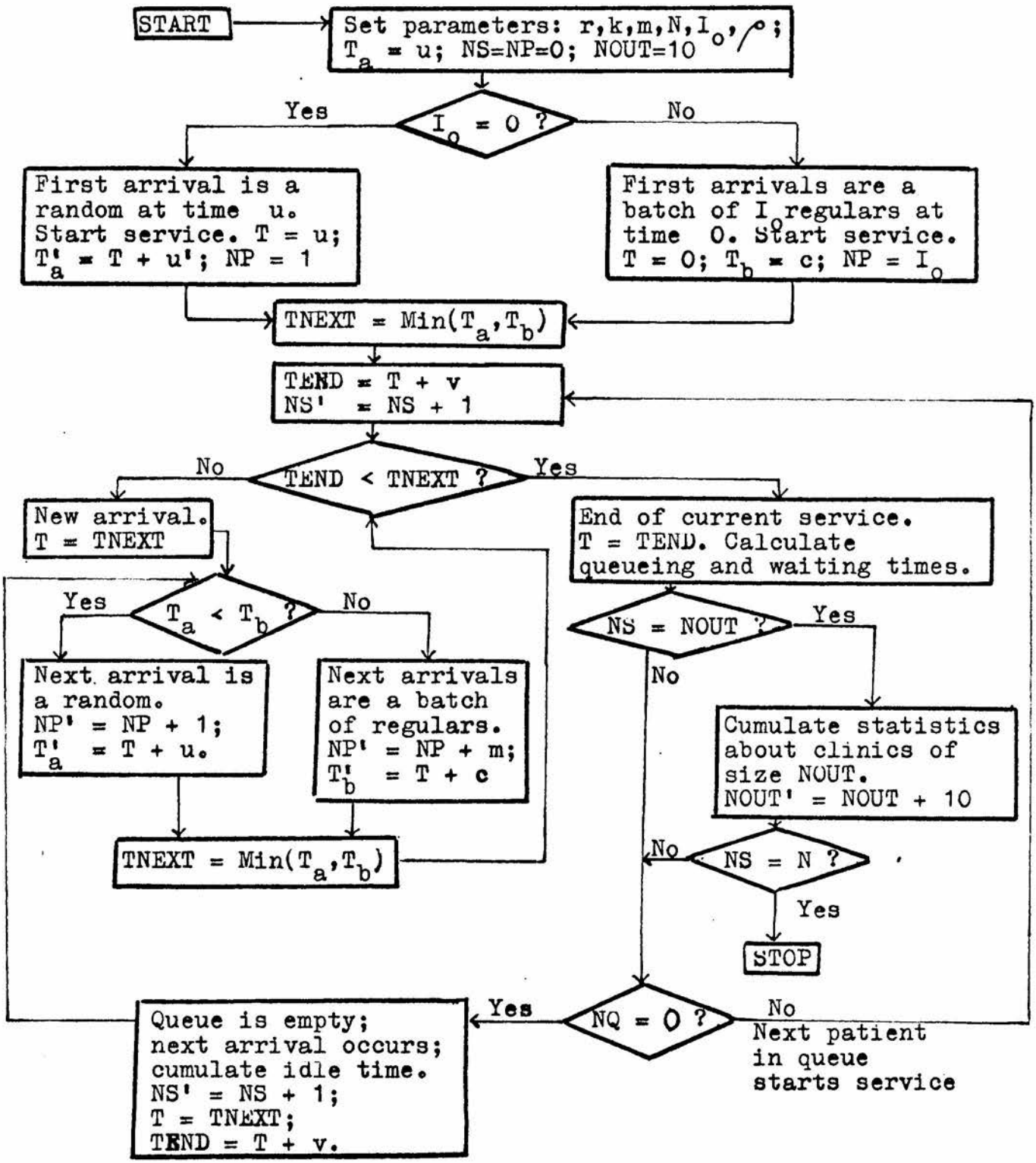


FIGURE 6.1 (continued)

Notation

T_a	=	Time of next arrival of a random.
T_b	=	Time of next arrival of a batch of m regulars.
u	=	Exponentially distributed pseudo-random variable (inter-arrival interval).
NS	=	Serial number of patient currently being served.
NP	=	Total number of patients who have arrived.
T	=	Current time.
TNEXT	=	Time of next arrival (of either class).
TEND	=	Time that present service is due to end.
NOUT	=	Next size of clinic for which information will be cumulated.
N	=	Largest size of clinic simulated.
v	=	Pseudo-random variable with E_k distribution (service time).
NQ	=	Number of patients queueing in line.
'		Indicates a change in a variable value.

At all times the computer "holds" in reserve the projected times of arrival of the next patients of each class, and determines which of these is to arrive first. Whenever the server is busy, the projected time of the completion of service is also stored. The question "TEND < TNEXT?" determines the nature of the next event, either an arrival or the end of the current service. When an arrival occurs, a new projected arrival time for the next patient(s) of the same class is set up; for appointment patients, the next batch of m people will be set to arrive in c time units; for non-appointment patients the time of the next single person's arrival is determined using a pseudo-random variable generator. When a service completion occurs, statistics on the queueing and waiting times of the patient are extracted. If there is a queue of patients, then the next in line begins service immediately; if, however, the server becomes idle, the program has to both cause the next arrival and initiate service before control is transferred as usual to the question "TEND < TNEXT?" A note is made of the idle-time incurred by the server.

To collect statistics about clinics of certain sizes, we extract the cumulated totals of idle time, and queueing and waiting times at various points in the program execution. As shown in Figure 6.1, this was done after the completion of service of the 10th, 20th, 30th, ... etc patients in each run. As the results for small clinics are thus derived from the first sections of simulations of larger clinics, we may see that the results for clinics of different sizes, with the same values for the other parameters, are not independent. (A comment is made on this in section 6.5.) Provision is made to store information about the two classes of patients separately. When the patient of serial number N (the largest size of clinic simulated)

has completed service, one run of the program is terminated. The program can be repeated an appropriate number of times for a given set of the defining parameters.

The machine used for the simulations was an I.B.M. 360/50, and the programming language was FORTRAN IV G. The size of the program was about 50K bytes.

6.3.1 Pseudo-Random Number Generator

The behaviour of the simulated system depends on sequences of random variables, such as the patient inter-arrival intervals and service times. To provide an input of such variables, the program makes use of a pseudo-random number generator, a device producing a sequence of numbers which has a random character. The members are produced by a mathematical recurrence relationship, and are thus not random in the true sense as successive numbers are not independent; if presented with a sequence of pseudo-random numbers, it would be possible in principle to "discover" the generation rule. However a suitable choice of generator will give a stream of numbers which may be regarded as random in practice; there is also the advantage that the numbers may easily be reproduced at some later stage.

Given an initial integer y_o, a sequence $\{y_i\}$ can be determined by a recurrence relation of the kind

$$y_{i+1} = (fy_i) \text{ Mod } m$$

where f and m are coprime integers. A scale transformation produces a sequence of numbers $\{x_i\}$ which are distributed in the interval $[0,1]$ Suitable choices of f and m will yield a sequence which passes many of the

tests of randomness, and has a long cycle. A further transformation

$$u_i = - a \ln(x_i)$$

gives variables u_i which are distributed exponentially with mean a. To generate pseudo-random variables with an E_k distribution, we simply have to add k pseudo-random exponential variables.

The comparison of results of different parameter values was felt to be more important than the numerical values of results for a particular set. Therefore it was decided to use the same streams (or sequences) of pseudo-random numbers to determine the inter-arrival intervals and service times in comparable simulations. This may have the effect of introducing some consistent sampling error into the results, but the sampling variation between groups of simulations will be reduced. As different random number streams are used in different runs with the same parameter settings, the absolute bias in any group of simulations will not be large if a sufficient number of runs are made with its parameter values.

6.4 Bias in Parameters caused by Program Stopping Rule

The program is designed to simulate clinics of up to size N, and each run is terminated after the completion of service of the Nth patient. The parameter r refers to the expected ratio of arrivals of the two patient classes during the main part of the clinic session's duration, but specifically excludes the starting instant. As we are now dealing with a finite queueing system, we would expect the actual ratio of the numbers of arrivals observed in a complete session to differ on average from r. Similar considerations apply to ρ. Any stopping rule seems to induce

a difference of this kind between the true (or "intended") and actual parameter values; however this particular rule does allow the comparison of results from different clinics of size N.

Let us again define α to be the rate of random arrivals. Then the expected number of random arrivals after the session has been running for a time t is αt, and the known number of regular arrivals is $I_o + \lfloor t/c \rfloor . m$, where $[x]$ is the largest integer not greater than x. We define $f(t)$ to be the expected number of all arrivals after a time t; this function is strictly monotonic increasing with gradient α, and step increments of size m at times c, 2c, 3c, ... We also have $f(0) = I_o$.

We first suppose that there is a value of t, t^* say, such that $f(t^*) = N$. (This value may not exist because of the jumps in the value of $f(t)$.) Then

$$\alpha t^* + I_o + \lfloor t^*/c \rfloor m = N \qquad (6.1)$$

We let $\lfloor t^*/c \rfloor = N^*$. Thus at $t = t^*$, the expected number of random arrivals is αt^*, and the exact number of regular arrivals is $I_o + N^*m$. We now let $\hat{r}$ be the ratio of the expected number of regular arrivals to the expected number of random arrivals during the whole clinic session. Then

$$\hat{r} = \frac{I_o + N^*m}{\alpha t^*} \qquad (6.2)$$

and

$$r = m/c\alpha \qquad (6.3)$$

As $\left[t^*/c\right] = N^*$, we must have that $t^*/c \geq N^*$, and so $1/t^* \leq 1/N^*t$. Thus

$$\hat{r} \leq \frac{I_o + N^*m}{\propto N^*c}$$

$$= \frac{m}{\propto c} + \frac{I_o}{\propto N^*c}$$

$$= r + I_o/\propto N^*c$$

$$= r + \varepsilon$$

We may regard ε as the maximum "bias" in r in differing from r. Alternatively, from (6.1), we may write

$$\varepsilon = I_o r/mN^*$$

$$= I_o r/(N - I_o - \propto t^*)$$

We secondly consider the case when there does not exist a value of t such that $f(t) = N$. Then there must be a value t^* such that $f(t^* - 0) < N$ and $f(t^* + 0) > N$. We let

$$f(t^* - 0) = I_o + \propto t^* + \left[(t^* - 0)/c\right] m = E_1$$

and

$$f(t^* + 0) = I_o + \propto t^* + \left[(t^* + 0)/c\right] m = E_2$$

The implication is that there is a batch arrival at time $t = t^*$, and so t^*/c is an integer, N^* say. Then

$$E_1 = I_o + \propto t^* + (N^* - 1)m$$

and

$$E_2 = I_o + \propto t^* + N^*m$$

We now suppose that $E_1 = N - \Delta$, where $\Delta > 0$. At time $t^* - o$ the expected number of random arrivals is $\propto t^*$, and the exact number of regular arrivals is $I_o + (N^* - 1)m$; all of these people will actually be served (as $E_1 < N$). At time $t^* + 0$, the expected number of random arrivals is still $\propto t^*$, but the number of regular arrivals is $I_o + N^*m$; only Δ of the m new arrivals will be served (in the simulation). Thus we have

$$\hat{r} = \frac{\text{(Expected total number of appointment patients served)}}{\text{(Expected total number of non-appointment patients served)}}$$

$$= (I_o + (N^* - 1)\,m + \Delta)/\propto t^*$$

Using the fact that $t^* = N^*c$, we have that

$$\hat{r} = \frac{N^*m + I_o - (m - \Delta)}{\propto N^*c}$$

$$= r + \frac{I_o - (m - \Delta)}{\propto N^*c}$$

$$= r + \varepsilon_1$$

We know that $0 < \Delta < m$ (because $E_1 = N - \Delta = E_2 - m$, and $E_2 > N$). Thus

$$\frac{(I_o - m)}{\propto N^*c} < \varepsilon_1 < \frac{I_o}{N^*c}$$

or

$$\frac{r}{N^*}\left(\frac{I_o}{m} - 1\right) < \varepsilon_1 < \frac{I_o r}{mN^*}$$

If $m > I_o$, ε_1 may be negative.

6.5 Choice of Parameter Values

In Chapter 5, it was necessary for practical reasons to restrict our attention to the solutions of models with a limited number of parameter values. The theoretical models involved four parameters m, k, r and ρ. For these simulations we have introduced two additional parameters, N and I_o. We might also consider the punctuality of the doctor at the start of a session; however as we saw in section 5.9.1, it is possible to consider the approximate effect of this variable by a suitable adjustment of I_o. For simplicity (by avoiding the inclusion of a further parameter) we will assume the doctor to be always punctual. With the simulation parameter space being six-dimensional, we must once again consider only a number of typical values.

An X-Ray department administration has direct control over the parameters I_o and m, and some indirect control over r, ρ and N. k is fixed for any given location and examination. As has been previously discussed, the X-Ray department is far from being able to devise an appointment scheme to its own advantage alone because of the interests of other units in the health service. It may be obliged to run a clinic of at least a certain size, and it may not be possible to adjust r or ρ beyond certain limits. Thus we are trying to maximise some function of the system over a severely restricted set of parameter values; in particular we may be forced to find a maximum over only a few allowable values.

To represent values of the parameters close to those which occur in practice, it was decided to simulate systems with the following:-

Batch size m = 1, 2, 3.

Initial number I_o = 1, 2, 3, 4.

Ratio r = 0.5, 1.0, 2.0.

Service distribution phase number k = 1, 2, 3, 4.

Size of clinic N = 10, 20, 30, 40, 50, 60.

Traffic intensity ρ = 1.0

Doctor punctuality p = 0

Not all combinations of the above were simulated. As is mentioned in section 6.3, information on clinics of different sizes were extracted during the same simulation runs; thus statistics about clinics of size 20 were obtained from the first thirds of runs simulating clinics of size 60. Such results are not independent, therefore, but it was felt that comparisons of clinics of the same size would be more important than between clinics of different sizes. By this means, a reduction by a factor of six is achieved in the computing effort.

Again to reflect common practice, it was decided that the simulations should be of clinics of length $2\frac{1}{2}$ hours or 150 minutes. (By length, we mean the expected time needed to serve all the patients; the clinic closing time will be somewhat later than $2\frac{1}{2}$ hours after the start because of idle-time.) However clinics of other lengths can be simply investigated by a scale transformation of the results. Within the program, the mean service time was taken as 5 minutes, and the results scaled appropriately for each clinic after the last run.

At the beginning of the program execution, it was necessary to determine the appointment interval c and the random arrival rate α in

terms of the specified parameters ρ, m, r and the mean service time b. We have that the rate of regular arrivals is m/c, so that the total rate of arrivals is $m/c + \alpha$. Also we have that $r = m/c\alpha$, so the total rate of arrivals can be written as

$$\frac{m}{c} \cdot \left(\frac{r + 1}{r} \right)$$

From the definition of the traffic intensity,

$$\rho = \frac{1}{b} \cdot \frac{cr}{m(r + 1)}$$

Thus in the program we use these formula in reverse to derive

$$c = \rho b m (r + 1)/r$$

and

$$\alpha = m / cr$$

After an initial period of experimentation it was decided that 25 runs would be necessary for each set of parameter values. With this number of runs, the estimates of the mean idle time per clinic and the mean patient queueing time had a standard error under 5% for all the parameter settings used. A file of dimension (25 x 60) was set up in the computer, containing generated variables distributed exponentially with mean one. One row (or part-row) of the file was used in each run to obtain the intervals between random arrivals; the filed variables were of course scaled by the factor $1/\alpha$. Similarly a file of the same dimension was set up for each value of k, and contained the service time variables scaled to have a mean of 5 (minutes).

6.6 Information Recorded during Simulation Runs

During the development phase of the program, it was necessary to print out a record of every event occurring in the model, that is all arrivals, departures, and times when the queue was empty, together with all the variables showing the state of the system. When the program appeared to be running correctly for all values of the parameters, the production stage was begun.

The information derived from the simulations concerned for the most part the idle time of the doctor and the queueing times of both classes of patients. The idle times observed during each run were recorded exactly, but the queueing times were formed into histograms with a grouping interval of $\frac{1}{2}$ minute and upper limit of one hour. From these, the mean and variance of all the variables were estimated. A combined histogram of the queueing times for all patients was formed, and the mean and variance calculated; estimates were made of the 75%, 90% and 95% percentiles of this distribution. Also recorded were the average number of patients of each type served in each size of clinic, and the average closing time (the end of the service of the Nth patient).

6.7 Results

Figure 6.2 shows the mean doctor idle time per clinic (y) and the mean patient queueing time (x) observed in four groups of simulations. The parameters r, k and m are constant in these groups, and have typical values. Different values of I_o lead to the four sets of results. During each run, information was cumulated after the completion of each ten patients; the results are shown after they have been scaled to represent comparable clinics each of length 150 minutes. The straight lines joining the points plotted

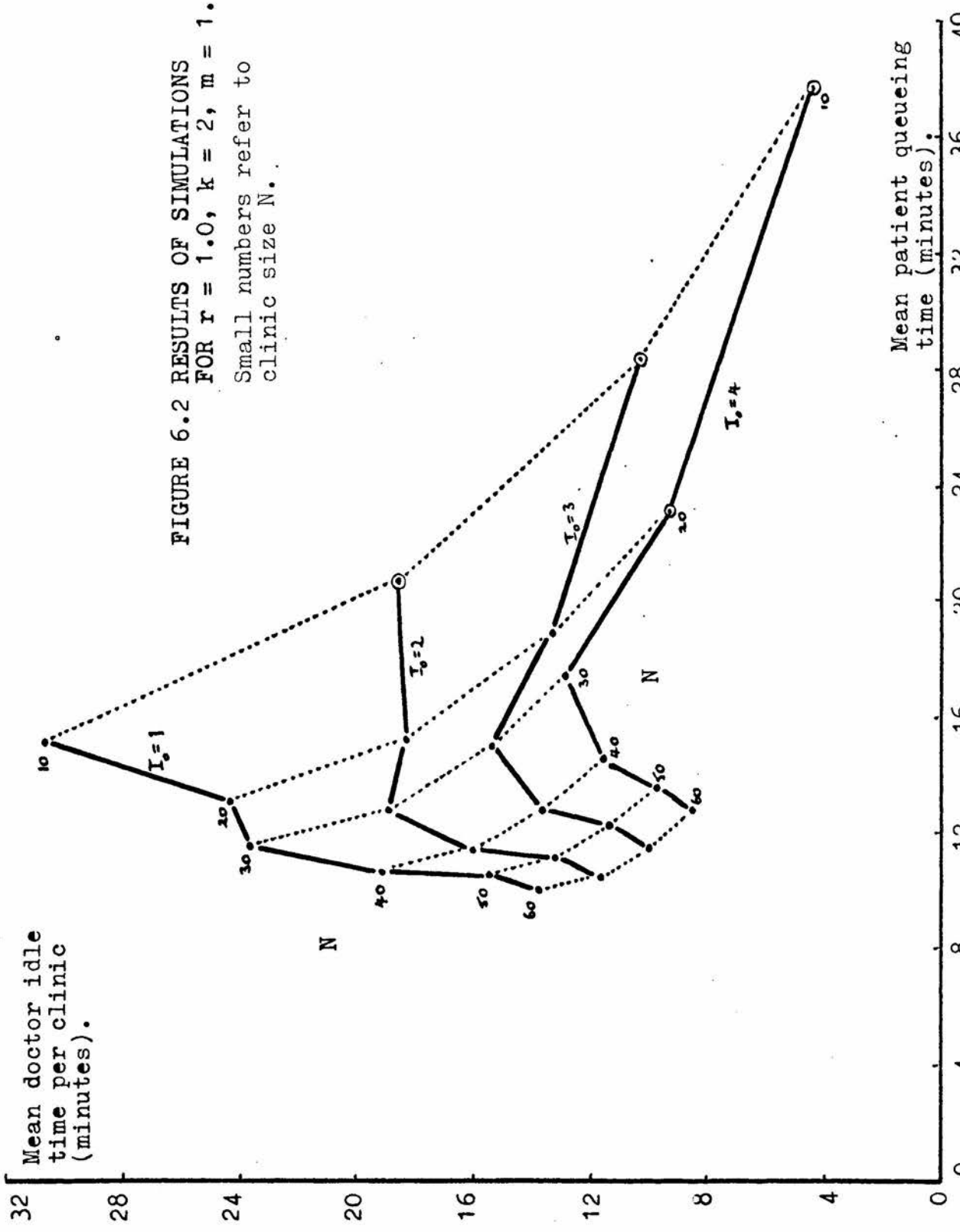


FIGURE 6.2 RESULTS OF SIMULATIONS FOR r = 1.0, k = 2, m = 1.

Small numbers refer to clinic size N.

are for clarity of presentation only: the "horizontal" lines joining points for clinics of different size might be regarded as a linear interpolation for clinics of intermediate sizes; however no physical meaning can be attached to the "vertical" lines joining clinics with different values of I_o (as only integer values of I_o are allowed).

As might be expected, higher values of I_o always give less idle time and more queueing times for clinics of the same size. In clinics with equal values of I_o, increasing the size usually decreases both idle and queueing times. For the larger values of I_o, however, we may see in Figure 6.2 that there is an area in the lower values of N where increases in the size are associated with increases in the idle time; considering the curve for $I_o = 4$ we may see this gives an intermediate higher value of the idle time before further increases in N are associated with the usual "south-westerly" movement of the curve. The effect of variation in the initial number I_o is less for clinics of larger size.

Where it was found that an appointment scheme did not meet the Ministry of Health's standard (that "not more than 25% of patients should wait more than 30 minutes"), the appropriate point on the Figure is circled. In Figure 6.2, it will be seen that in only small clinics with a higher number of patients present at the start did this apply, and in fact this was the case for most of the simulations.

Figures 6.3, 6.4, 6.5, and 6.6 show the same cross-section of the response surface for other parameter values. They each have the same characteristics as Figure 6.2 but with different emphases reflecting their particular parameter settings. In Figure 6.3 only the value of k differs

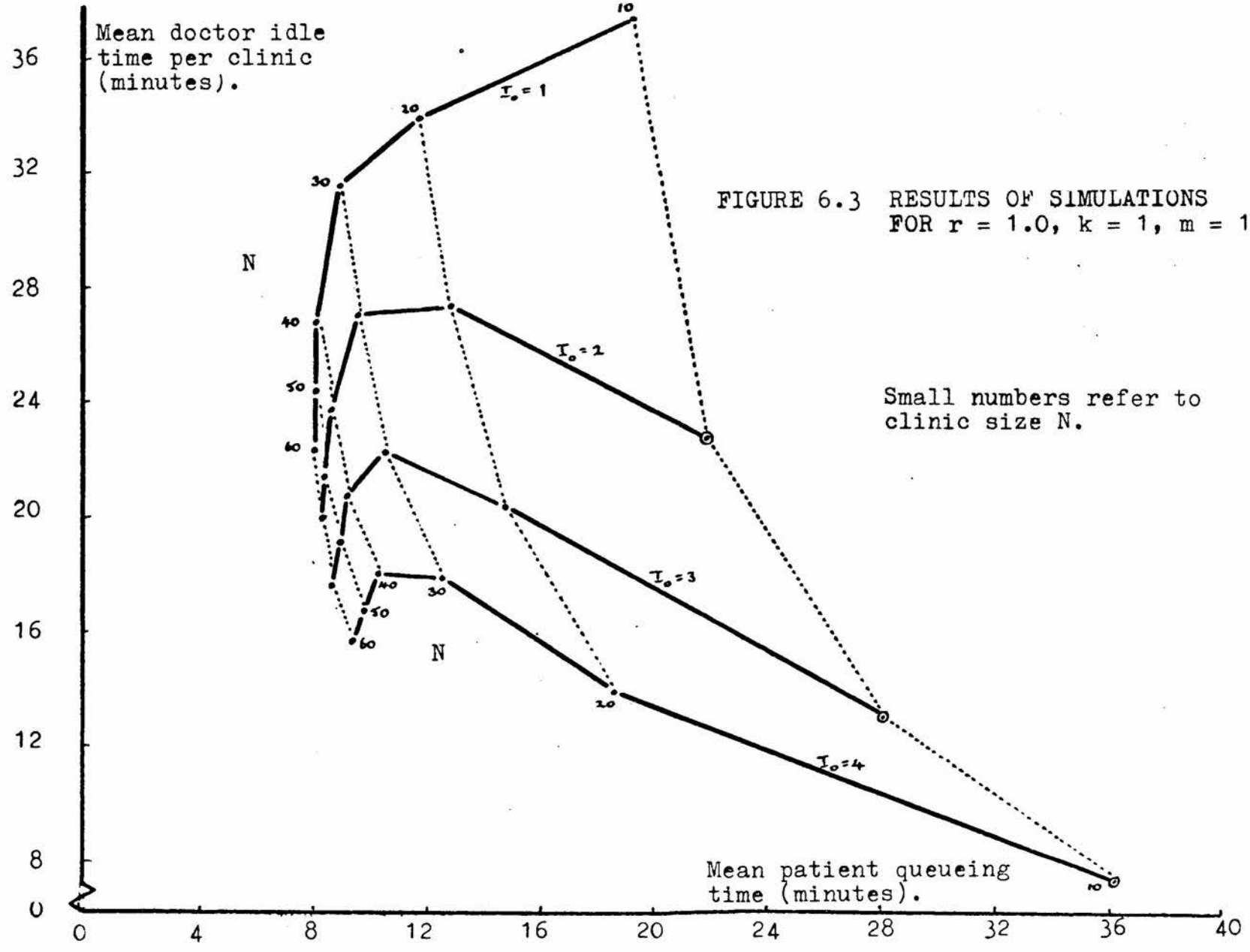


FIGURE 6.3 RESULTS OF SIMULATIONS FOR r = 1.0, k = 1, m = 1.

Small numbers refer to clinic size N.

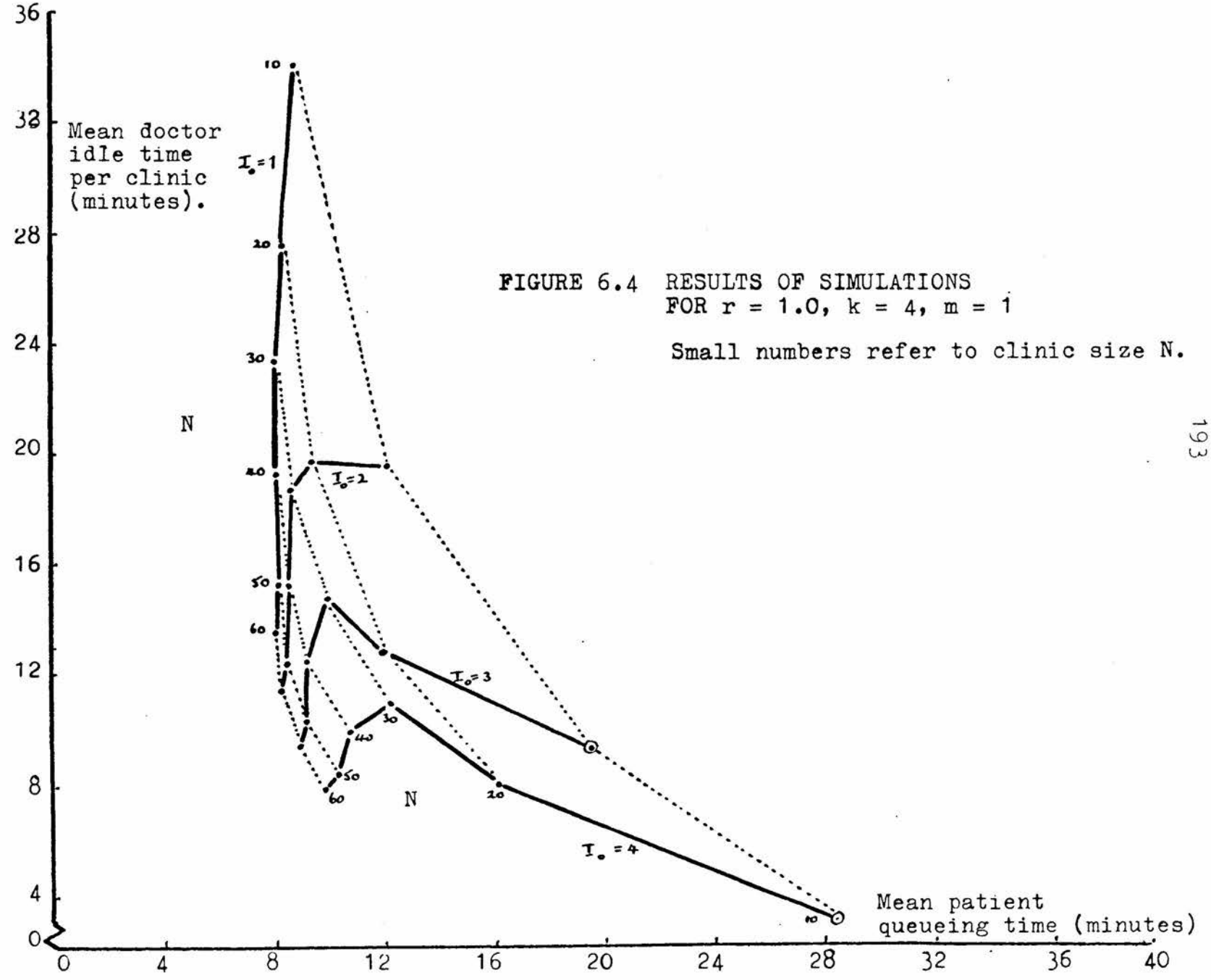


FIGURE 6.4 RESULTS OF SIMULATIONS FOR r = 1.0, k = 4, m = 1

Small numbers refer to clinic size N.

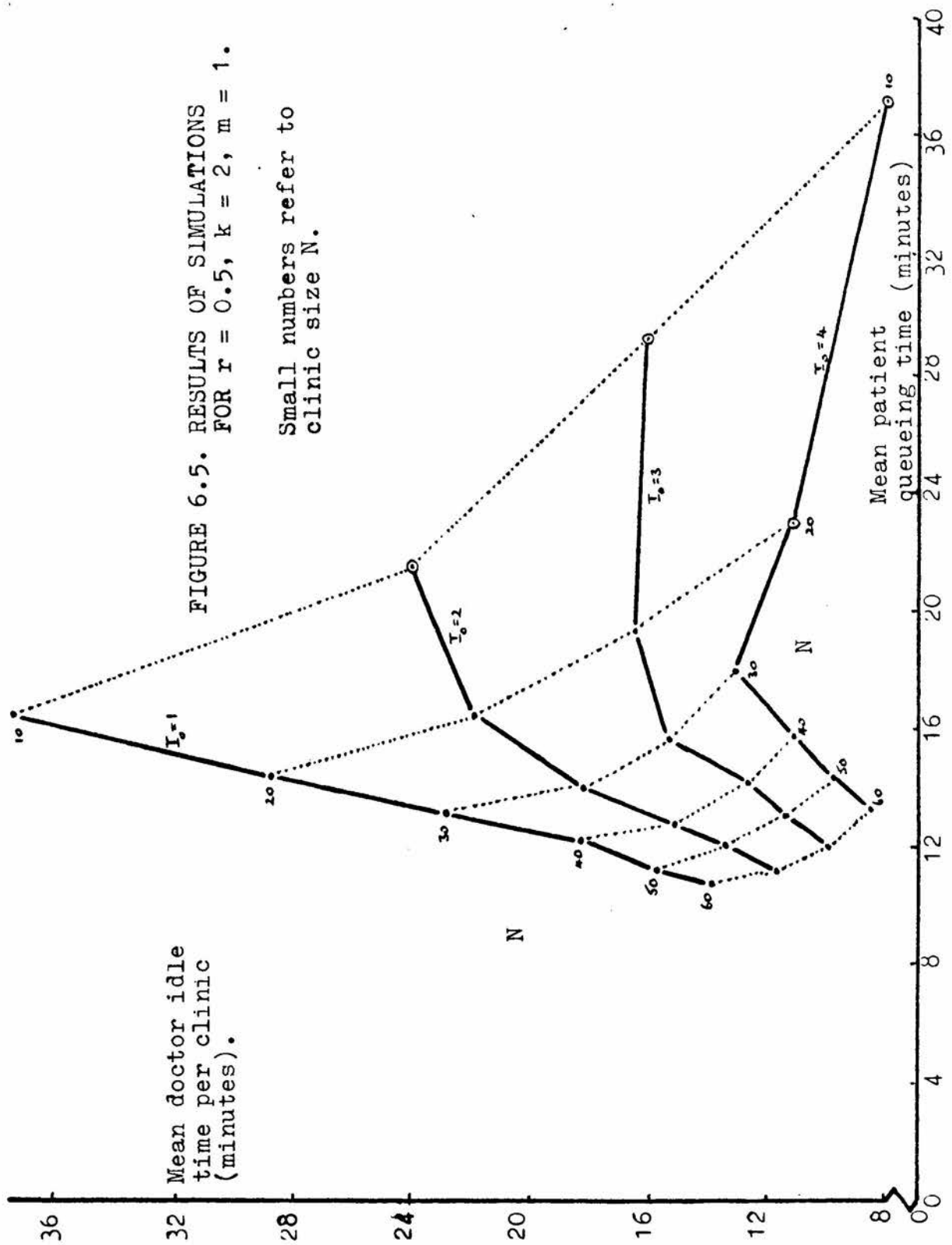


FIGURE 6.5. RESULTS OF SIMULATIONS FOR $r = 0.5$, $k = 2$, $m = 1$.

Small numbers refer to clinic size N.

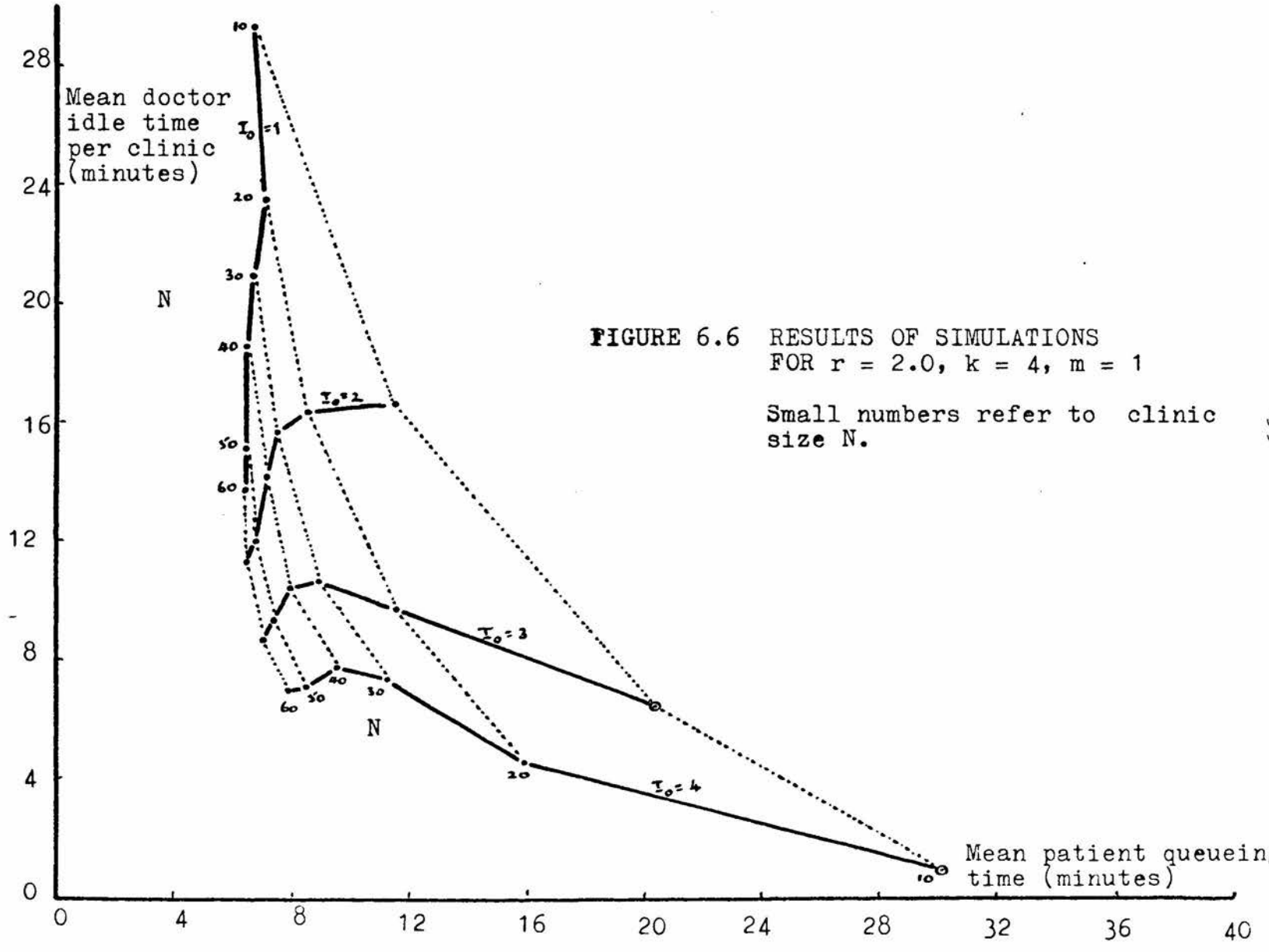


FIGURE 6.6 RESULTS OF SIMULATIONS FOR r = 2.0, k = 4, m = 1

Small numbers refer to clinic size N.

from Figure 6.2; k = 1 (as opposed to k = 2) implies a more variable service time distribution (in fact the exponential), and we may note increases in x and y for comparable clinics. In contrast Figure 6.4 shows the surface for k = 4. This is a much less variable distribution (it may be remembered that as $k \rightarrow \infty$, the service distribution becomes regular, or constant), and reductions in both x and y are to be seen, some of which are considerable. As an example, the results for r = 1.0, m = 1, I_o = 3, N = 20 are (x, y) =(14.8, 20.3) when k = 1 and (x , y) = (11.9, 12.9) when k = 4. In Figure 6.4 we may observe that with low values of I_o and larger clinics, the mean patient queueing time becomes almost constant, with idle time reducing with increasing clinic size.

In Figure 6.5, the value of r has been reduced to 0.5 from 1.0 in Figure 6.2. This means we are now considering clinics with a lower proportion of patients arriving with appointments, and we may see that comparable clinics are not so efficient, in terms of x and y, as previously. The whole response surface cross-section has moved (and distorted) in a direction generally "north-east", meaning higher (and less desirable) mean idle and queueing times. In other words, it would seem (from these observations at least) that increasing the value of r is desirable from both staff and patient considerations. In contrast, in Figure 6.6 which shows the results for r = 2 (more patients by appointment) and k = 4 (a less variable service time distribution), we have results representing very efficient appointment schedules. Either x or y, or both have been substantially reduced at all points as compared to Figure 6.2; we may also note the mean queueing times for I_o = 1 which are almost constant (as in Figure 6.4). These results demonstrate a

tendency which is evident in the other figures to a lesser extent, that for low values of I_o it is mainly y which varies with N, and with higher values of I_o it becomes x that shows the greater change. However, it should once again be pointed out that as the results have been scaled to give clinics of constant length, we are not comparing clinics of different sizes having the same service time distribution; with the same parameter values (including I_o), we may only compare clinics of different sizes with the same service time distribution by using unscaled results. This is mentioned later in this section.

Figures 6.7, 6.8 and 6.9 show the effect of changing r while keeping m and I_o fixed, each figure corresponding to a different value of k. The most useful comparisons to be made from these results are between clinics of the same size. It may be seen that in almost all cases an increase in r results in a decrease in x or y, or both. The only exceptions in Figure 6.7 are firstly for k = 1, N = 10, where a slight increase in x is seen, secondly k = 2, N = 60 where there is a slight increase in y, and lastly for a few medium values of N there are small increases in y between two values of r only. Some of these exceptions are undoubtedly to be explained by sampling error; however all these increases in the "wrong" directions are very small, and are always associated with much larger decreases in the other variable. In Figure 6.9, where k = 4, we may once again observe a mean queueing time which is almost constant for larger clinics; most striking is the curve for r = 0.5, for which x shows a very small decrease from 10.9 minutes when N = 20 to 9.1 minutes for N = 60.

We may observe in each of these figures that when N is small, increases in r result mainly in a decrease in y, with little effect on x ; when N is

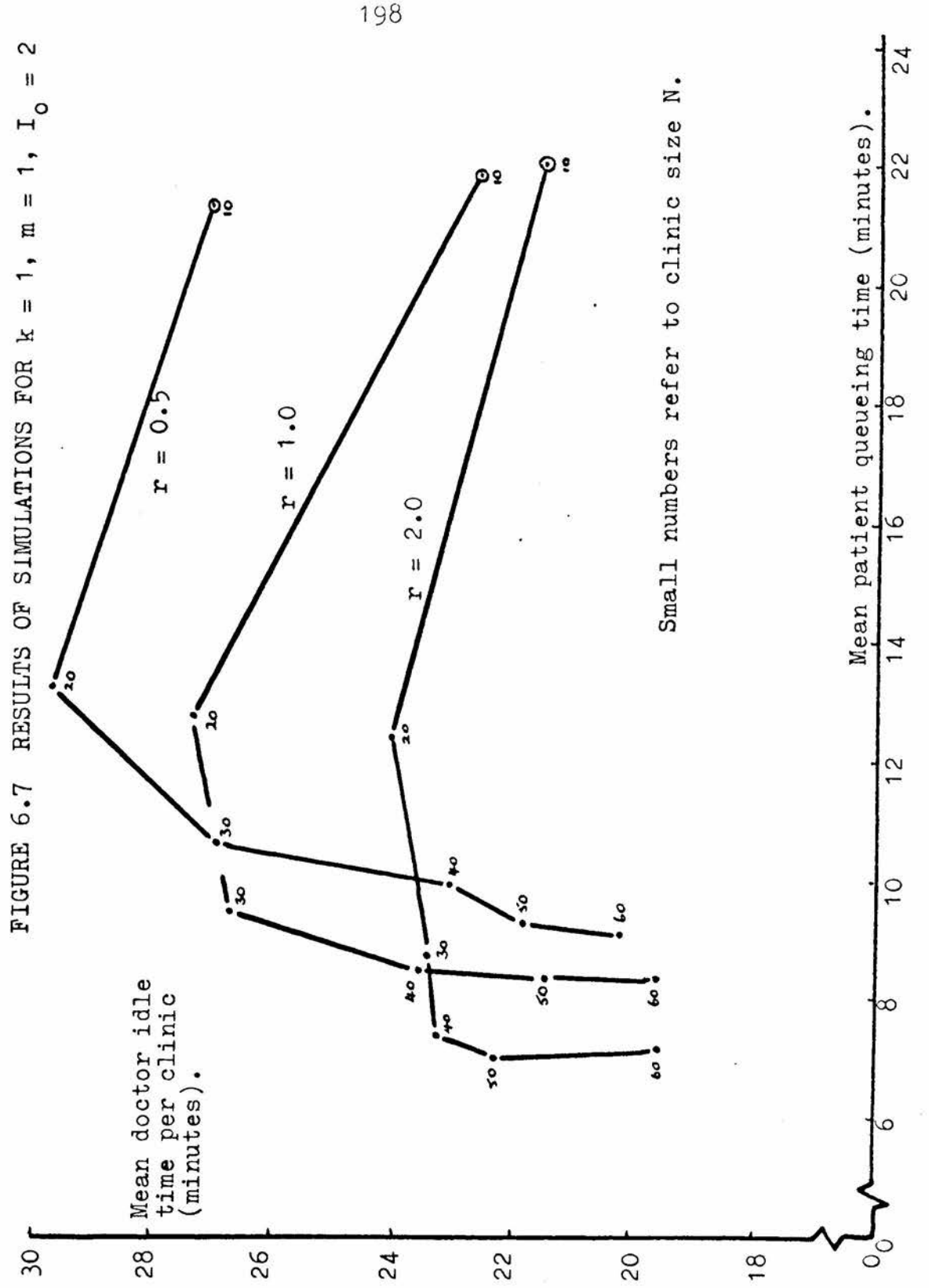


FIGURE 6.7 RESULTS OF SIMULATIONS FOR $k = 1$, $m = 1$, $I_o = 2$

FIGURE 6.8 RESULTS OF SIMULATIONS FOR k = 2, m = 1, I_o = 2.

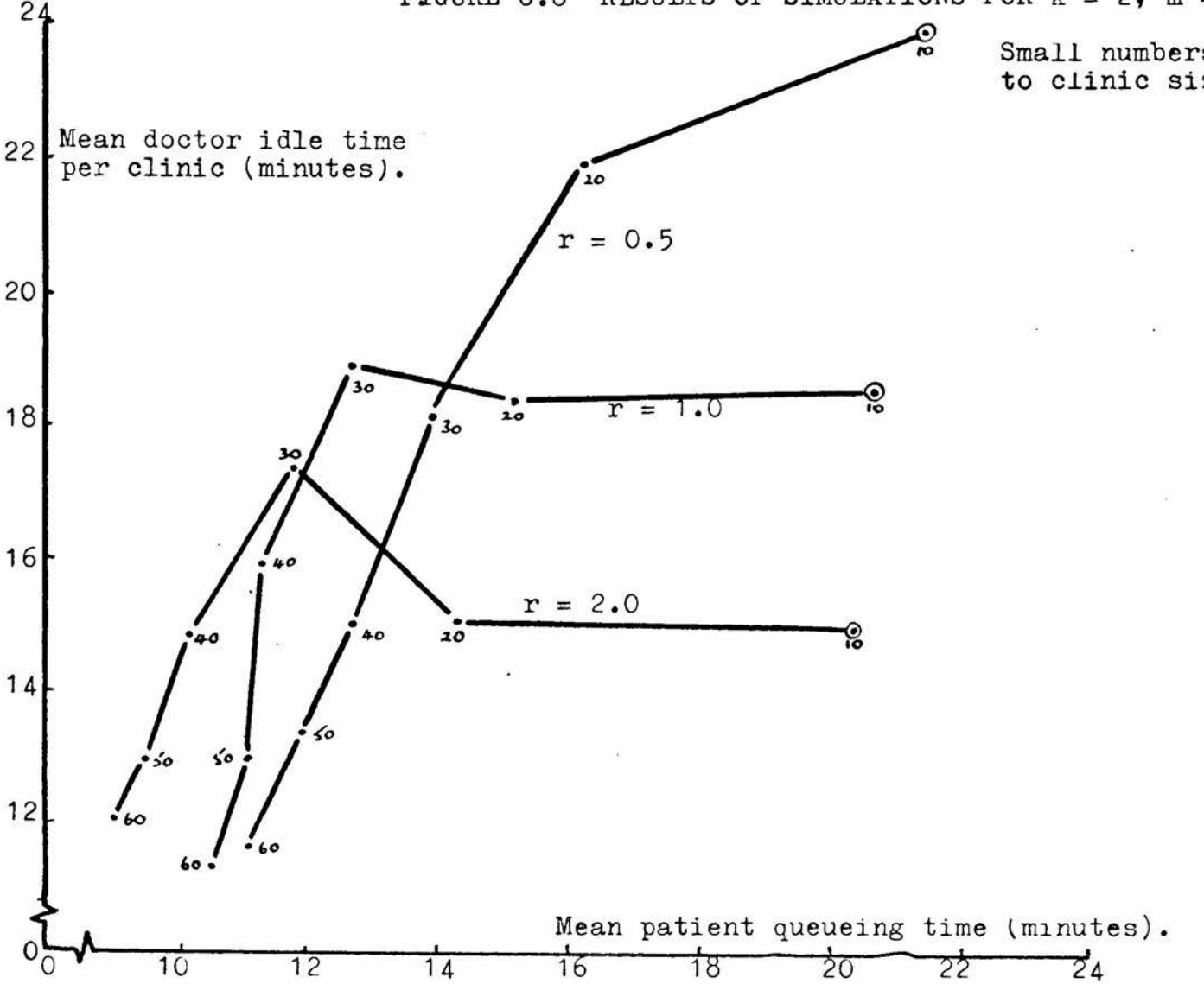
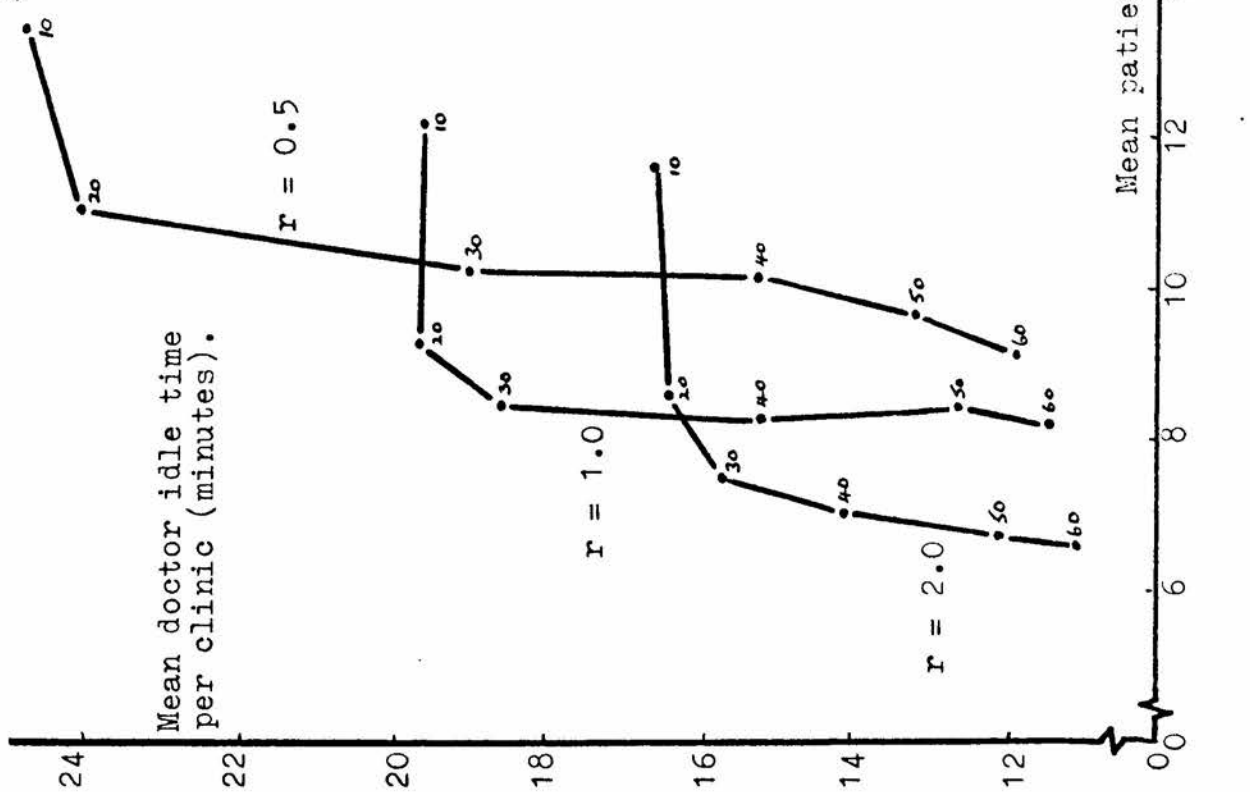


FIGURE 6.9 RESULTS OF SIMULATIONS FOR $k = 4$, $m = 1$, $I_o = 2$.

Small numbers refer to clinic size N.



large it is x that becomes much more sensitive to changes in r. For intermediate values of N, both x and y decrease when r increases, and in general a "south-west" movement is the result. We thus reach the important conclusion that if the proportion of appointment patients is increased, in small clinics it is the doctor who benefits by a reduction in his expected idle-time, and for larger clinics it is to the patients' advantage because there is a decrease in the expected queueing time.

Figures 6.10, 6.11 and 6.12 attempt to demonstrate the effect of the batch size m. Although the other parameters may be held constant, if the initial number I_o is also constant, it would appear that a rather meaningless comparison is obtained when the batch size is varied. In practice the initial group is treated just as a particular group arrival at the start of the session; any increase in m would presumably be accompanied by an increase in I_o. For simplicity I_o is taken as equal to m when making this comparison. A refinement of the model is therefore temporarily lost here, but an appointment system of m patients at times 0, c, 2c, ... does not seem unreasonable or unrealistic.

Referring to the figures 6.10, 6.11 and 6.12, we may see that for large clinics the effect of m is small; if anything, an increase in m seems to be against the interests of the patients (x is increased) and in favour of the doctor (small decrease in y), but the effect is small. When we consider medium-sized clinics the differences become more pronounced. For example in Figure 6.10, when N = 30, we may see a "south-easterly" movement from (x , y) = (11.5, 23.7) when m = 1 to (x , y) = (14.4, 21.1) when m = 3. In general, successive increases in m yield progressively smaller reductions in y and larger increases in x (there are, of course, exceptions

FIGURE 6.10 RESULTS OF SIMULATIONS FOR r = 1.0, k = 2, m = I_o.

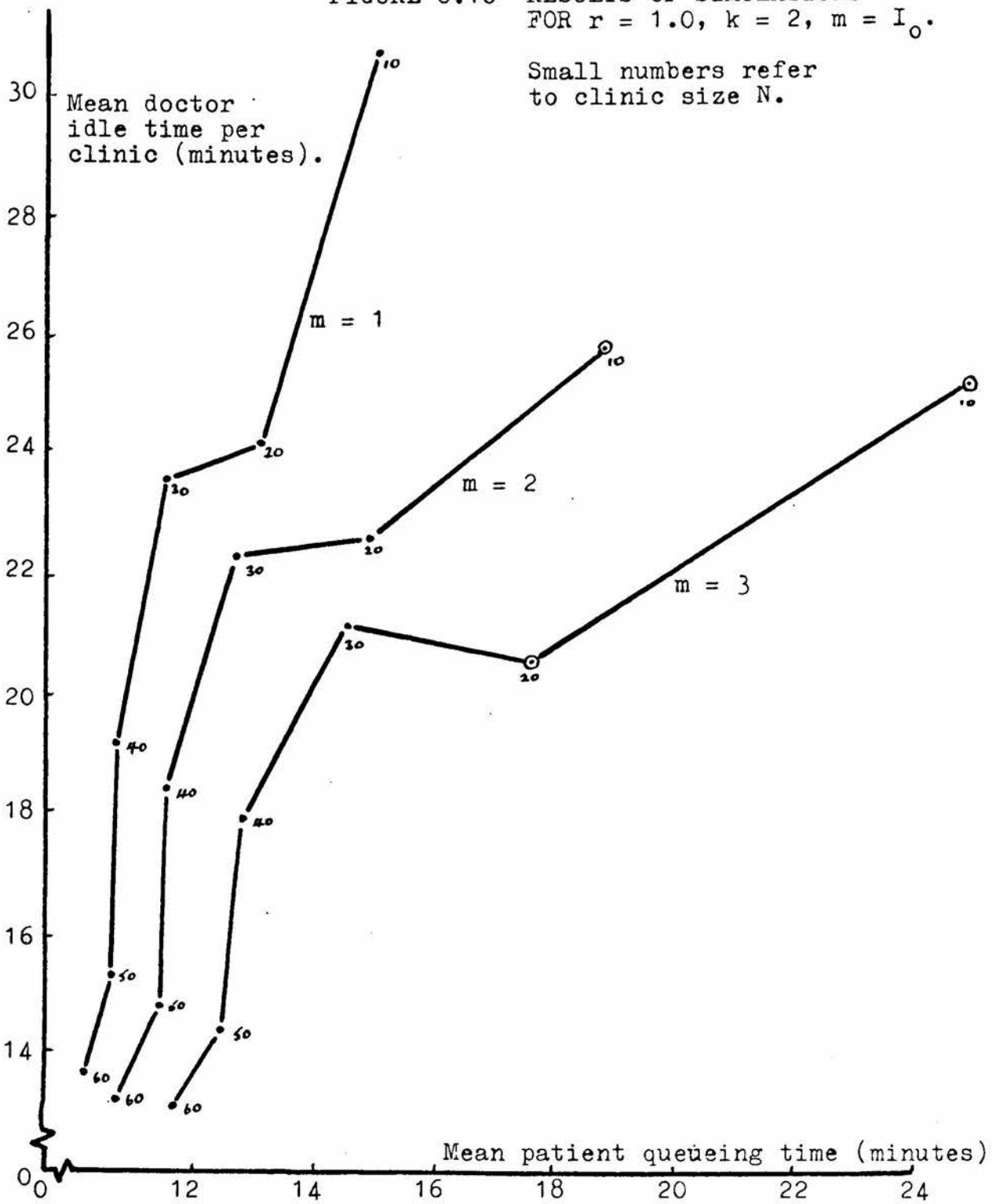


FIGURE 6.11 RESULTS OF SIMULATIONS FOR r = 0.5, k = 1, m = I_o.

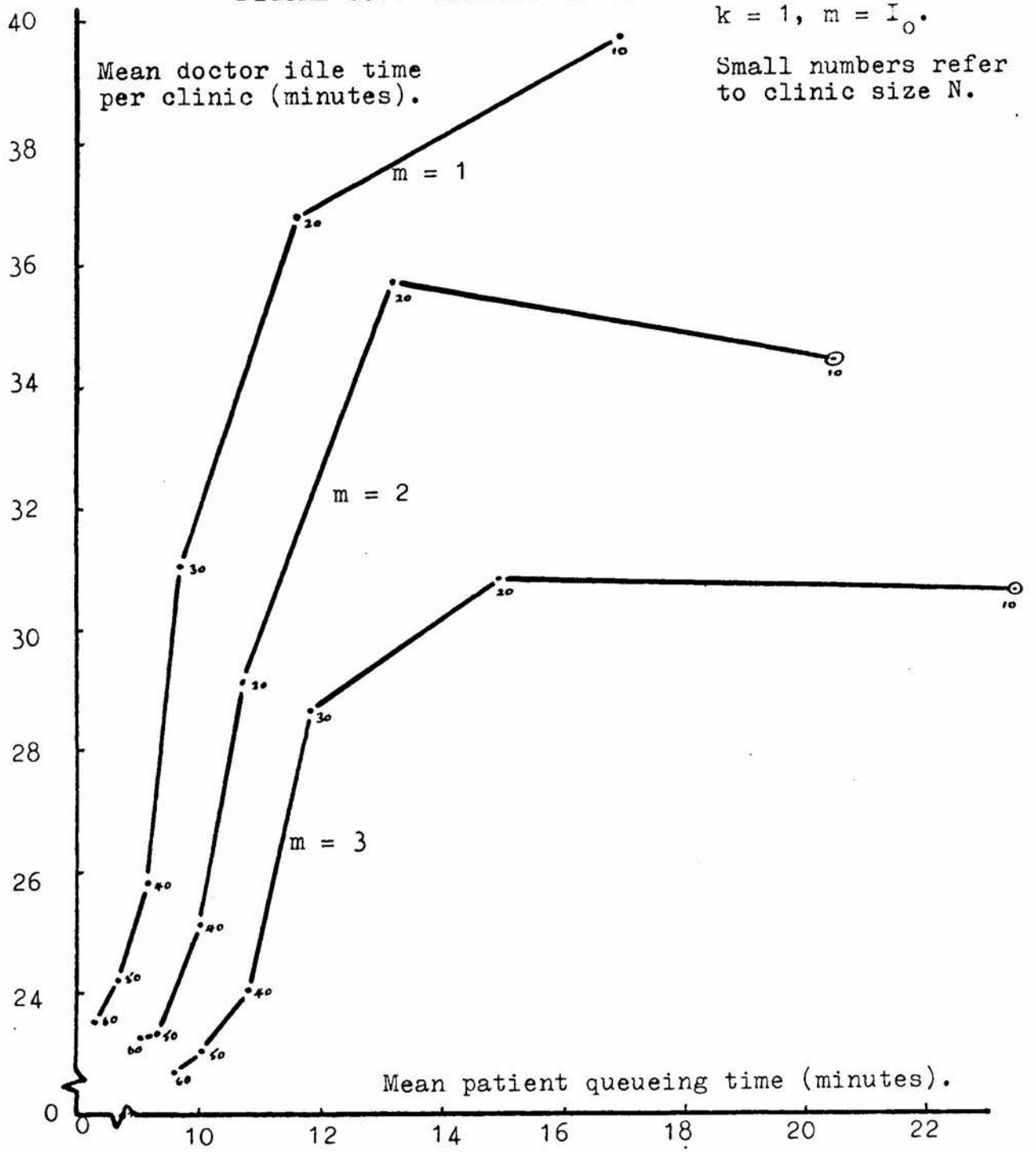
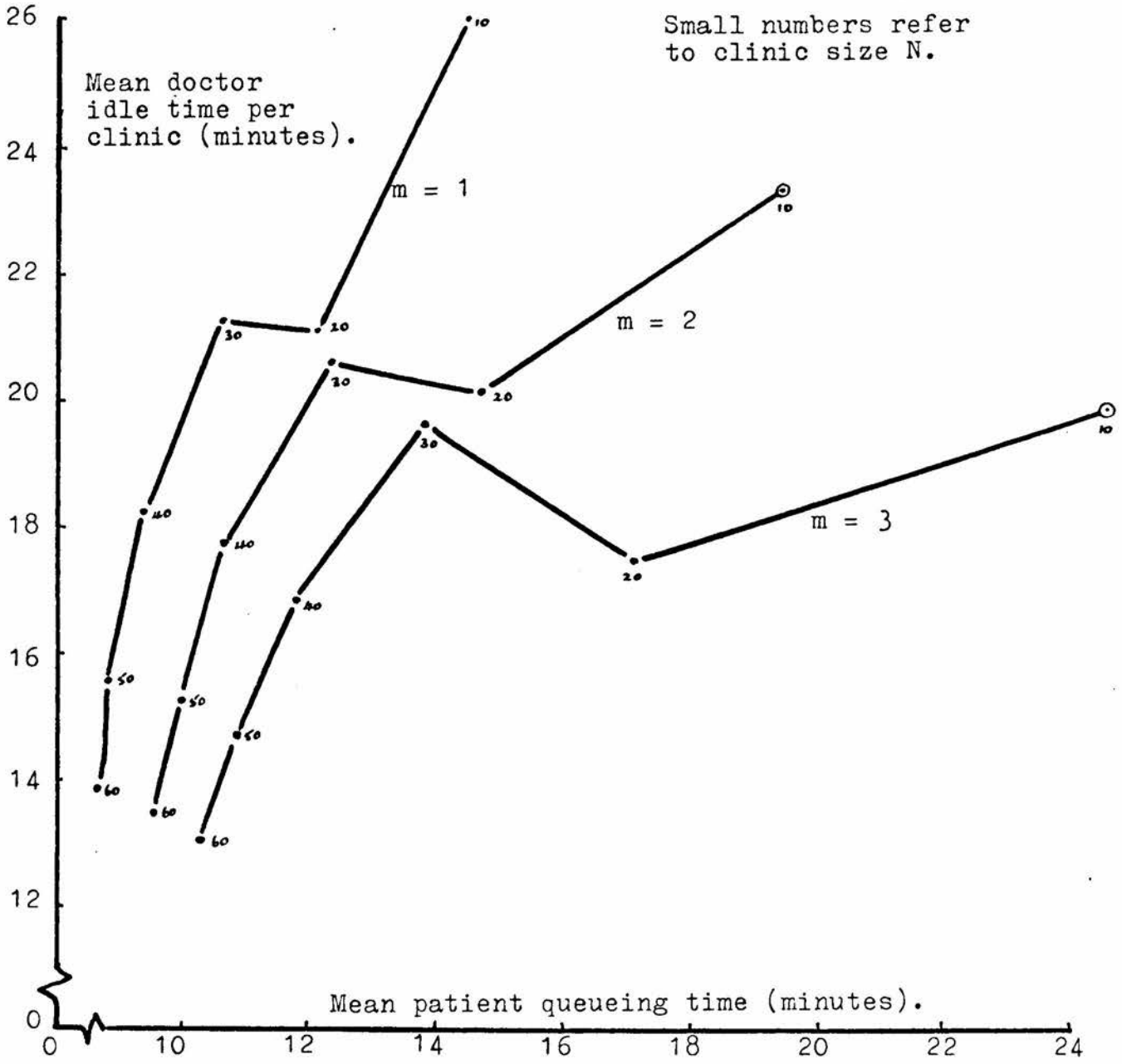


FIGURE 6.12 RESULTS OF SIMULATIONS FOR r = 2.0, k = 2, m = I_o.



to this). Therefore although values of m other than one may be justifiable because the idle time is reduced, we must remember that an increase in x implies a longer wait for each patient in the session. Also m must not be made so large that only small reductions in y are bought at the expense of large increases in x . This is particularly true for small clinics where x may increase substantially between m = 1 and m = 2; an example of this is in Figure 6.12. For small clinics it is very hard (on this basis) to justify a choice of batch size other than one. However clinics with a variable service time distribution (low values of k in this case) do seem to profit sometimes from block bookings (see for example Figure 6.11, N = 20).

Figures 6.13 to 6.18 make further comparisons of the results using the total queueing time by all patients in a clinic session. These results have an important alternative interpretation. Figures 6.2 to 6.11 had as abscissa the mean queueing time per patient x , scaled so as to give clinics of the same total length; the scaling factor was thus inversely proportional to the size of the clinic, N. In Figures 6.13 onwards, we are now plotting, as abscissa, a variable N times that used previously; so we have now an abscissa proportional to the mean queueing time per patient. In symbols the new abscissa is

$$z = \left(x \times \frac{h}{N}\right) \times N$$

where h is a constant of proportionality. (In fact h = 30: the clinics are designed to have a length of 150 minutes using a service time mean of 5 minutes). We may interpret z either as the total queueing time of all patients in clinics of constant length (as plotted in Figures 6.13 to 6.18), or, "cancelling" the N's, as hx which is simply proportional to

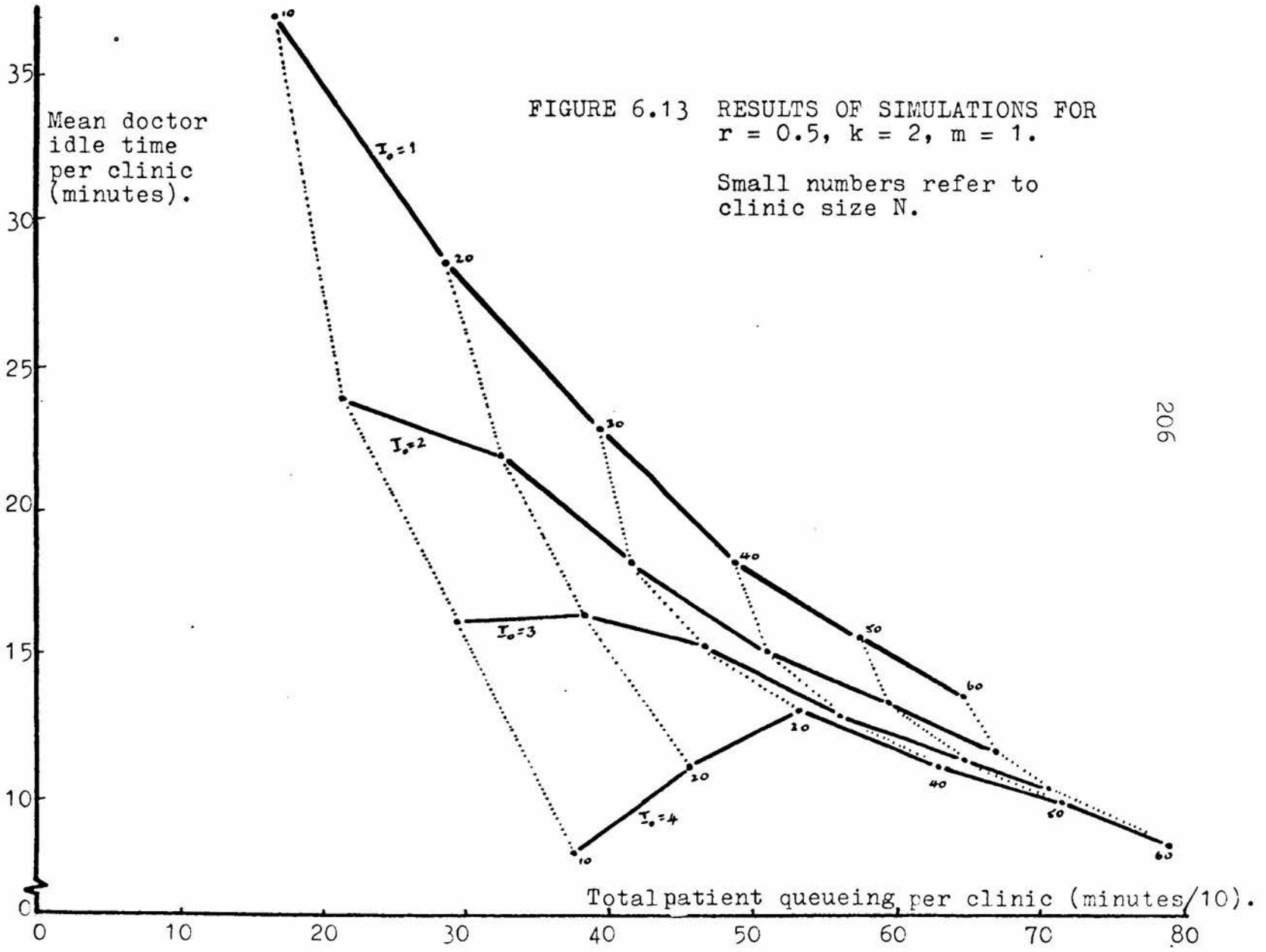


FIGURE 6.13 RESULTS OF SIMULATIONS FOR $r = 0.5$, $k = 2$, $m = 1$.

Small numbers refer to clinic size N.

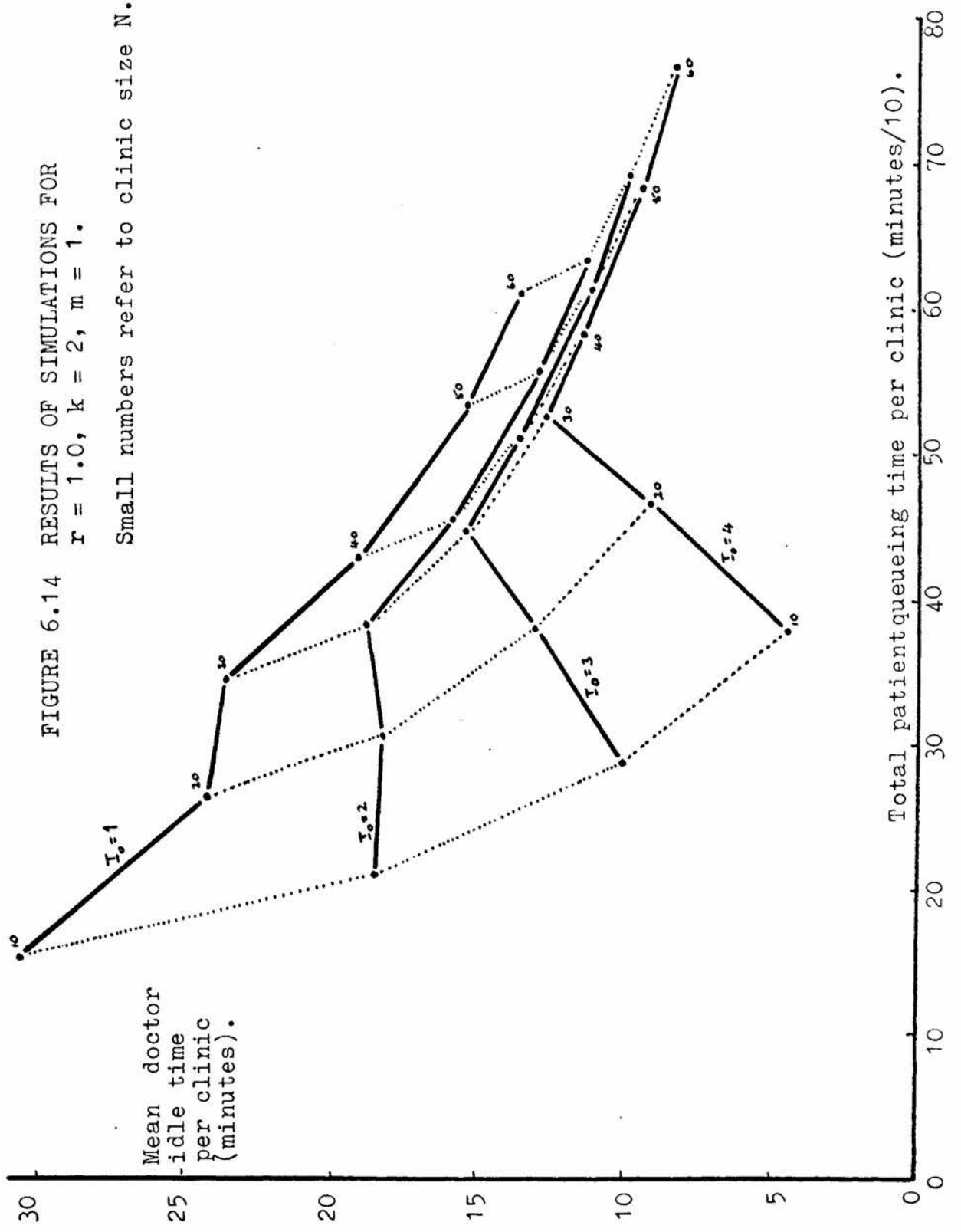


FIGURE 6.14 RESULTS OF SIMULATIONS FOR
r = 1.0, k = 2, m = 1.

Small numbers refer to clinic size N.

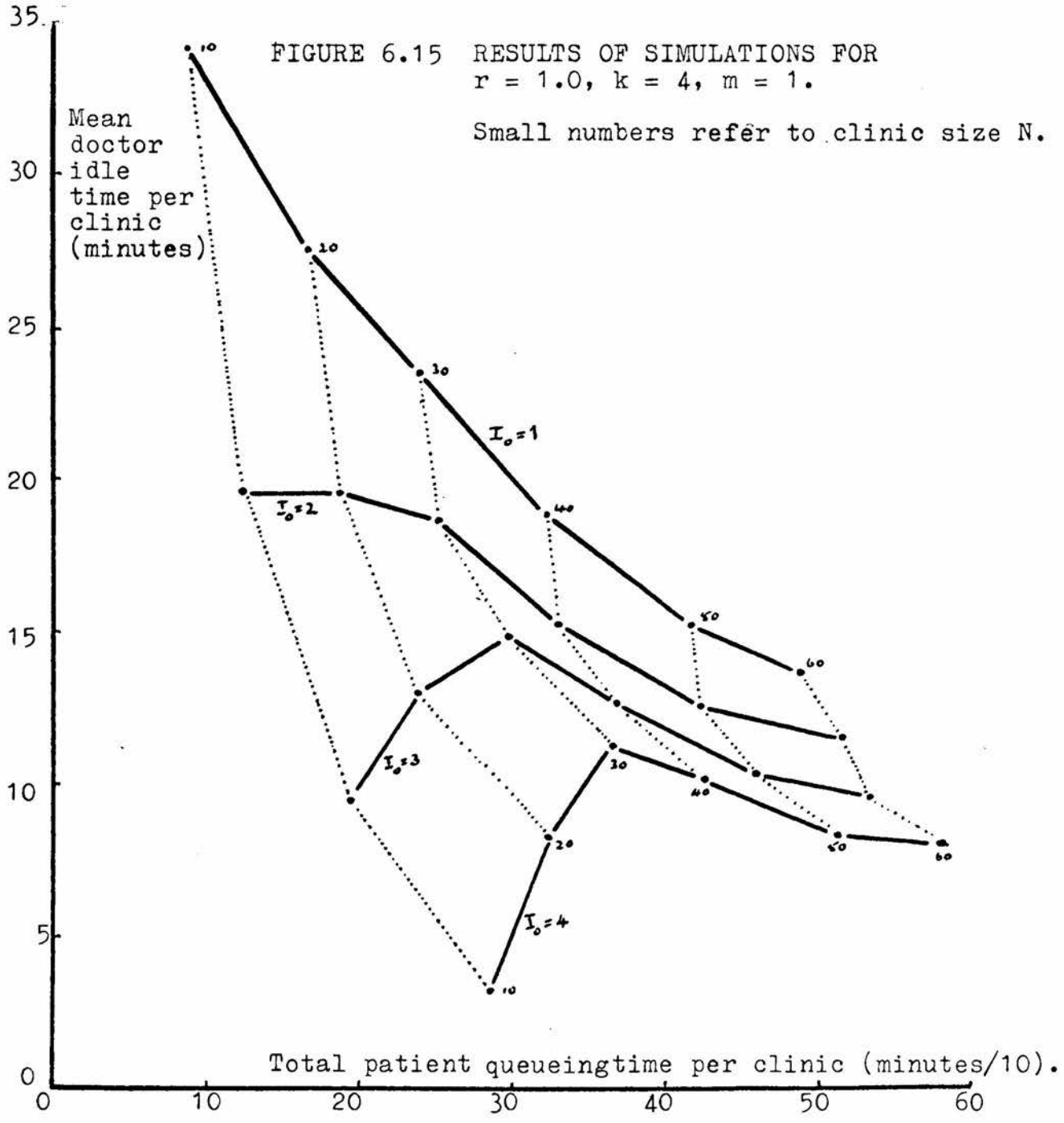


FIGURE 6.15 RESULTS OF SIMULATIONS FOR
$r = 1.0$, $k = 4$, $m = 1$.

Small numbers refer to clinic size N.

FIGURE 6.16 RESULTS OF SIMULATIONS FOR
$k = 2$, $m = 1$, $I_o = 1$.

Small numbers refer to clinic size N.

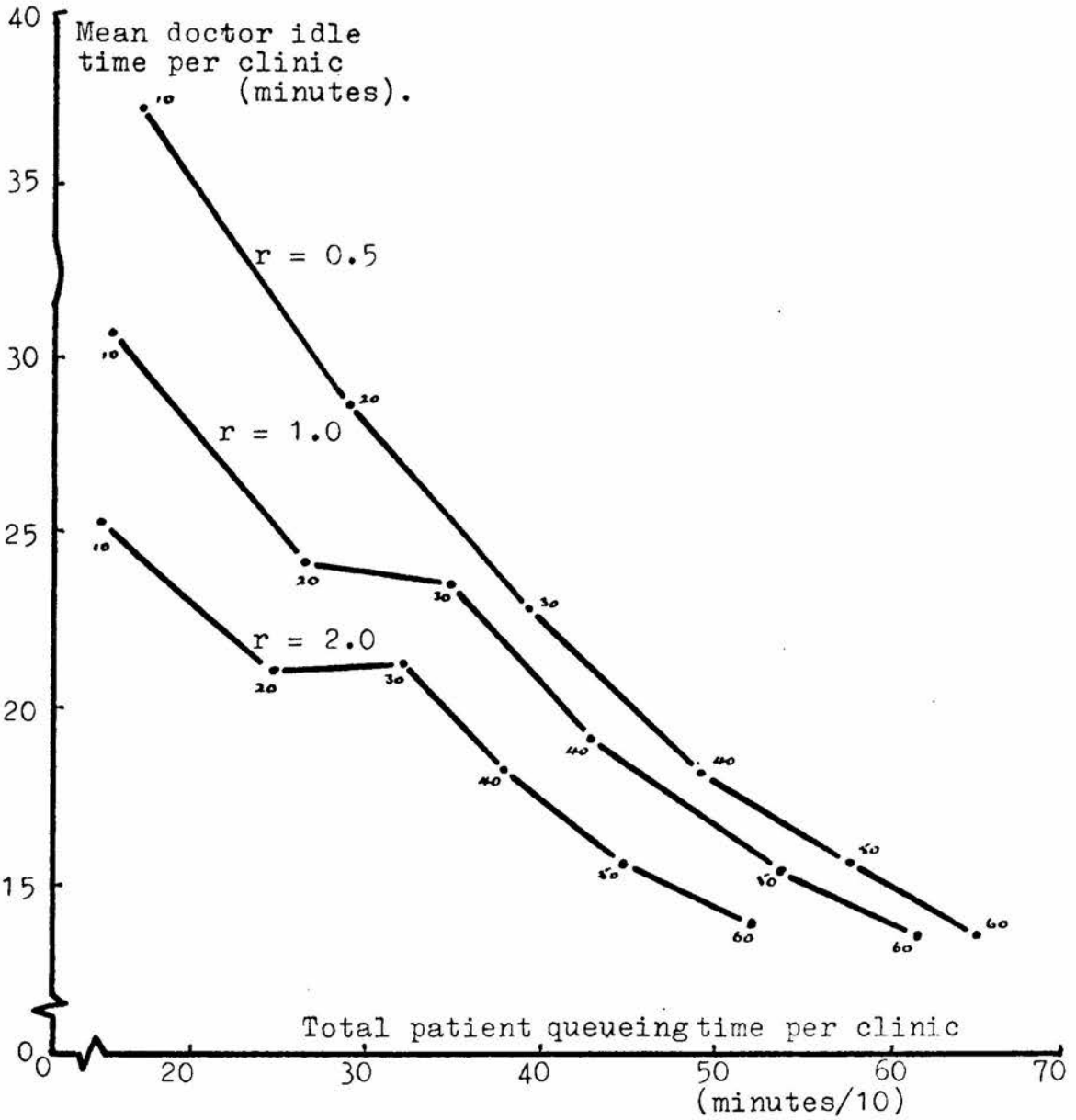


FIGURE 6.17 RESULTS OF SIMULATIONS FOR
$k = 2$, $m = 1$, $I_o = 4$.

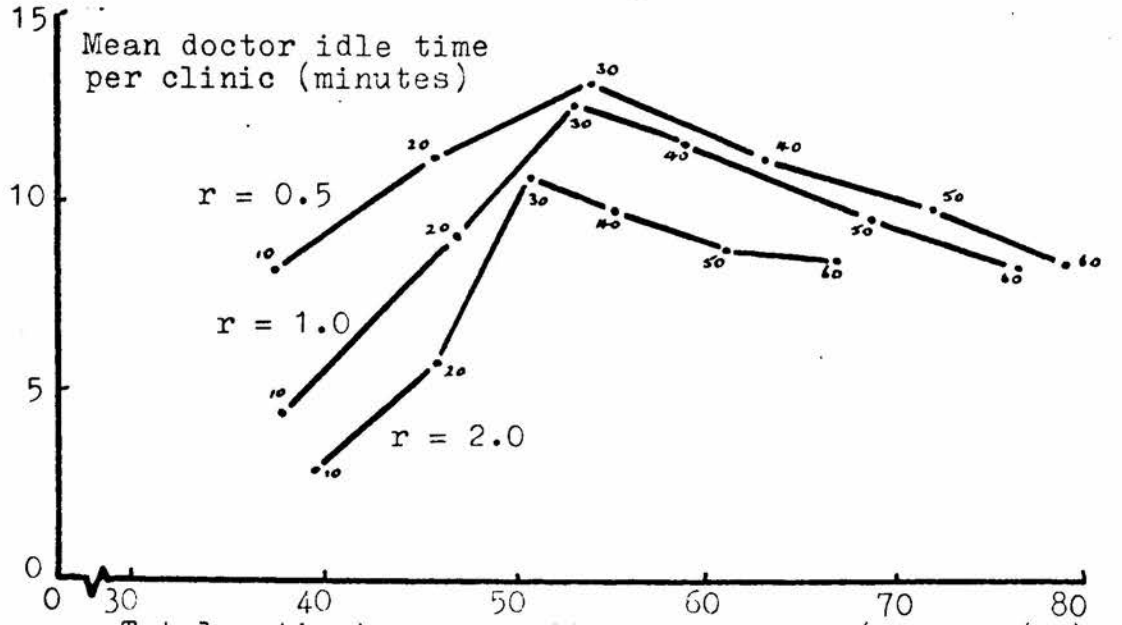
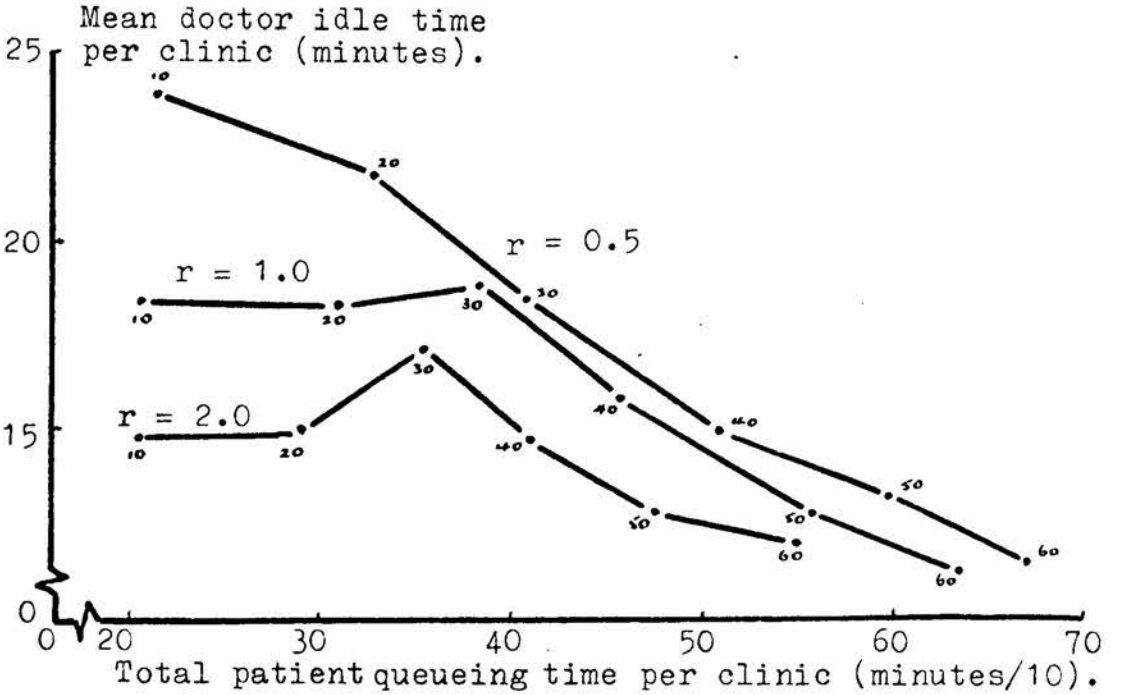


FIGURE 6.18 RESULTS OF SIMULATIONS FOR
$k = 2$, $m = 1$, $I_o = 2$.



the original mean queueing time. Thus the alternative interpretation is that of comparison of clinics of different sizes with the same service time mean, in terms of the mean queueing time per patient. To obtain actual figures for this mean, it is simply necessary to divide the plotted abscissa by h (which is 30 for a service time mean of 5 minutes). The ordinate values have remained unchanged, and to obtain the equivalent comparison of y values for clinics with the same service mean, the y values plotted in Figures 6.13 onwards must be scaled by a factor N/30.

Figures 6.13 to 6.18 give an alternative demonstration of some of the principles noted for the earlier results. We may again see that the effect of I_o is small in large clinics, and that the more efficient systems are obtained when r and k are large. Figures 6.16, 6.17 and 6.18 in particular illustrate the advantages of increasing the value of r.

If it were possible to specify the relative values of the times wasted by a doctor and the times wasted by all his patients, then, in principle, figures like 6.13, 6.14 and 6.15 could be used to select optimum values of I_o, given the other parameters. If it was decided, say, that a doctor's time was worth λ times that of one of his patients (on average), then we would wish to choose the appointment schedule that had a minimum value of the function $\lambda y + z$. We thus wish to find the values of I_o which gives a clinic point lying on the line $\lambda y + z = c$ having the smallest possible value of c. As an example, if $\lambda = 10$, $r = 0.5$, $k = 2$, $m = 1$ (see Figure 6.13), we find that the total value of times wasted is a minimum for $I_o = 3$ when $N = 10$, $I_o = 2$ when $N = 20$, 30, 40 or 50, and $I_o = 1$ when $N = 60$. With the same value of λ, and $r = 1.0$, $k = 4$, $m = 1$, then the best systems are obtained when $I_o = 3$ for $N = 10$ or 20, $I_o = 2$ for $N = 30$ or more.

6.8 Simulations using an Age-Differential Service Time Distribution

The above results are from simulations which assumed the same service time distribution for all patients; as we saw in section 4.6.3 a more refined model would be obtained by assuming a distribution which is a function of age, mobility and origin of the patient. A number of further simulations were performed to do this, and also to investigate the effects of dealing with work in sessions of patients from particular sources, as outlined in section 4.6.4. It was decided to first simulate clinics dealing with a majority of outpatients, and then clinics devoted principally to inpatient work, both performing the chest X-Ray.

To decide on the constitution of work for each group of simulations, the data collected from the department archives were used. Referring to Table 4.3, we may see that the total of 397 chest examinations included in the sample was made up of 226 cases from the hospital wards (inpatients), 27 from Casualty, and 144 from all the other sources (outpatients). We will now make, in this example, the following assumptions: that 75% of the Casualty patients arrive during normal departmental working hours (we are now discussing a department with integrated casualty services); that 95% of the inpatients have appointments, and that 5% of the outpatients have appointments. (Other values for these proportions could be used for a given hospital.) We thus have the following arrival pattern, in terms of relative numbers of patients:-

	Patient Source		
	Wards	Casualty	Outpatients
Arriving by appointment	214	0	7
Arriving without appointment (random)	12	20	137

We now wish to divide these patients into sessions of two types, nominally "inpatient" and "outpatient". We first put the 214 appointment ward cases into the "inpatient" sessions, and the 137 random outpatients into the "outpatient" sessions. The rest have to be divided in some appropriate proportions. Table 4.10 gives the mean service time for the chest examination for various patient groups; from these data we may derive the means for inpatients and outpatients examined on both machines as 3.18 and 2.60 minutes respectively. To a reasonable approximation, the combined lengths of all clinics of the "inpatient" and "outpatient" types will be in the ratio (214 x 3.18) to (137 x 2.60); simplified, this becomes 65 : 35. Now we need simply to divide the remaining patients (12 random inpatients, 20 casualty patients, and 7 appointment outpatients) in this ratio to obtain a final allocation to the two clinic types as:-

	Session Type		
Patient Group	"Inpatient"	"Outpatient"	Total
Appointment inpatients	214	0	214
Random outpatients	0	137	137
Random inpatients	8	4	12
Appointment outpatients	4	3	7
Casualty patients	18	9	27
Total	244	153	397

Within the "inpatient" session, we thus have a value for r, the ratio of appointment to random arrivals, of 218/26 = 8.4; for the "outpatient" session the value is 3/150 = 0.02. Simulations were carried out using these values of r, and using the distributions of patient ages given in Figures 4.5 to 4.8. After each arrival of either a random or appointment

patient, a pseudo-random number was used to determine the patient group. Then the age group of the patient was determined using the empirical distributions of Figures 4.5, 4.6 and 4.7 and a second pseudo-random number. Finally by using the data of Table 4.9, the appropriate service time distribution mean was calculated and used to generate the actual service time in the simulation run.*

The results of this set of simulations are given in Figures 6.19 to 6.24, again for just a few typical parameter values. Figures 6.19 to 6.22 relate to the "outpatient" clinics, and Figures 6.23 and 6.24 give a comparison for the "inpatient" sessions. The overall service time mean for the "outpatient" clinics is 2.62 minutes, and once again the results have been scaled to give all comparable clinics the same length; the size of clinic with scale factor of one was 30, and so the results refer to a clinic length of (30 x 2.62) = 78.6 minutes. The equivalent figures for the inpatient sessions are a service time mean of 3.12 minutes and a clinic length of 93.6 minutes. Clinics of different lengths can be investigated by a linear transformation of the results.

We may see that the idle time is very much less in the inpatient session for all parameter values than in the comparable outpatient clinics. In clinics of small size and high initial number, the queueing time is actually slightly more in the inpatient sessions; this difference is always overshadowed by a much larger change in the idle time. For larger

---

*It was assumed here that the value of k was fixed for the same examination on all classes of patient. This may not be true, but insufficient data were available to fit a different distribution for each class. More detailed study along these lines would be needed to do this, and also to estimate accurately other parameters, such as, for example, the percentage of outpatients having appointments.

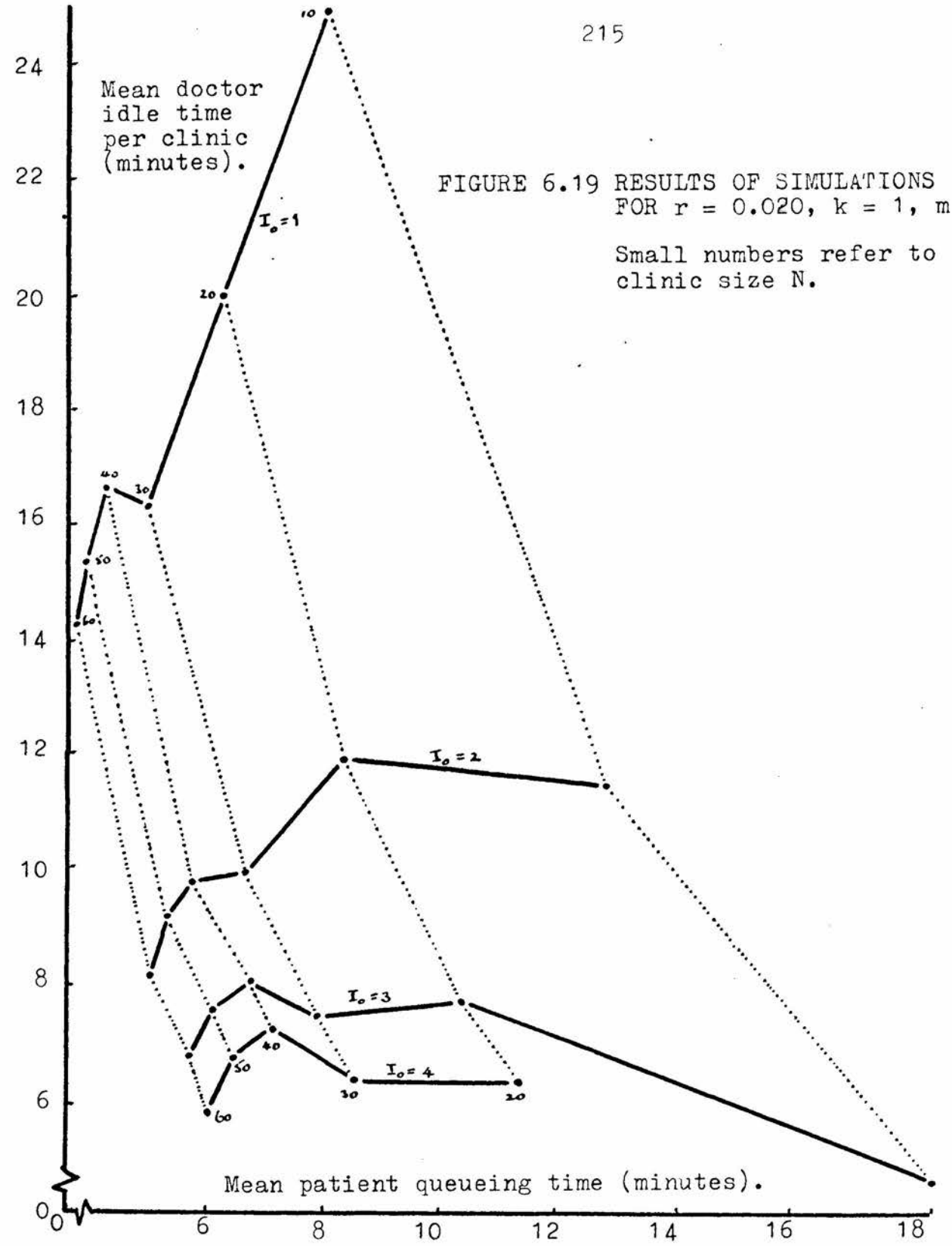


FIGURE 6.19 RESULTS OF SIMULATIONS FOR r = 0.020, k = 1, m = 1.

Small numbers refer to clinic size N.

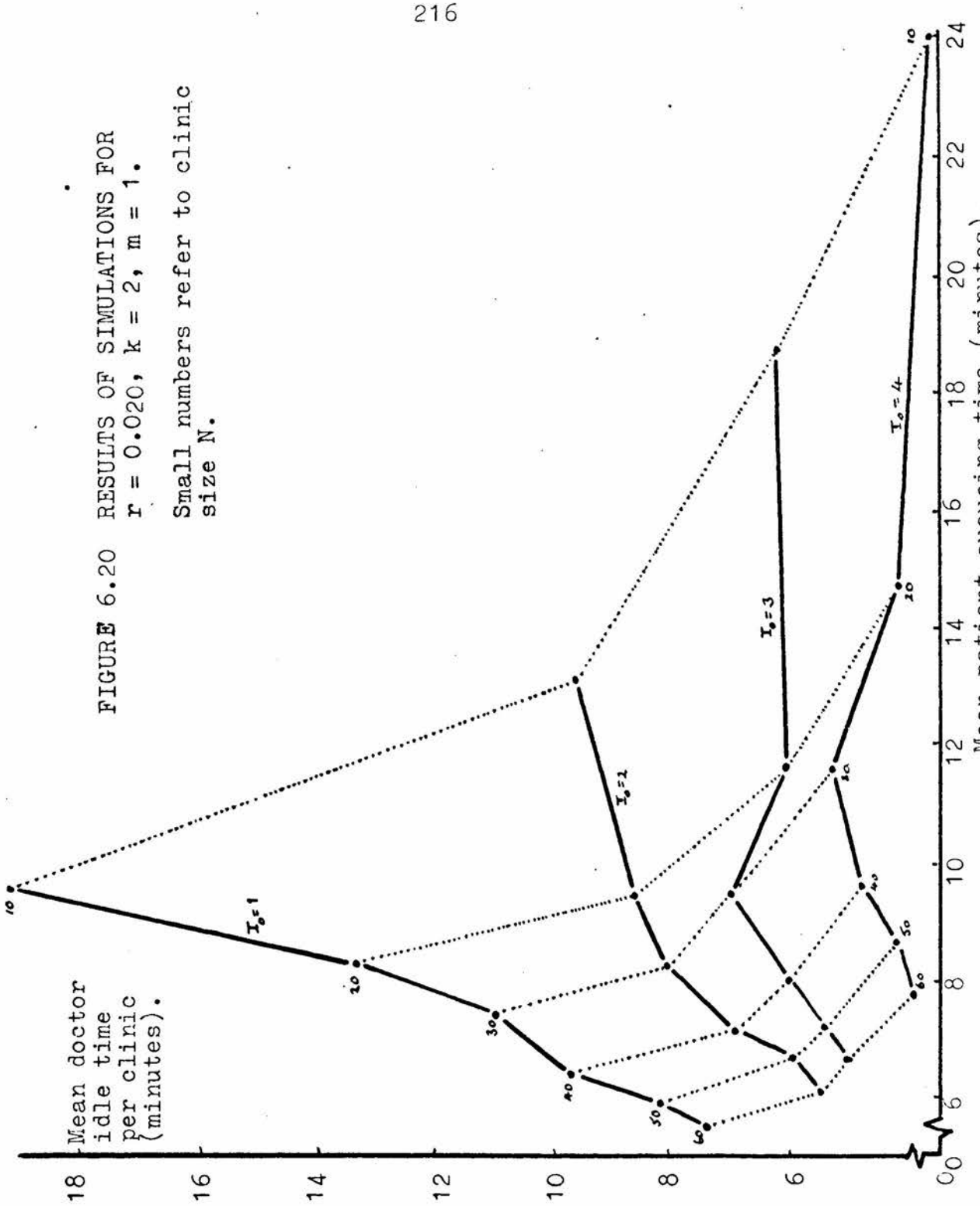
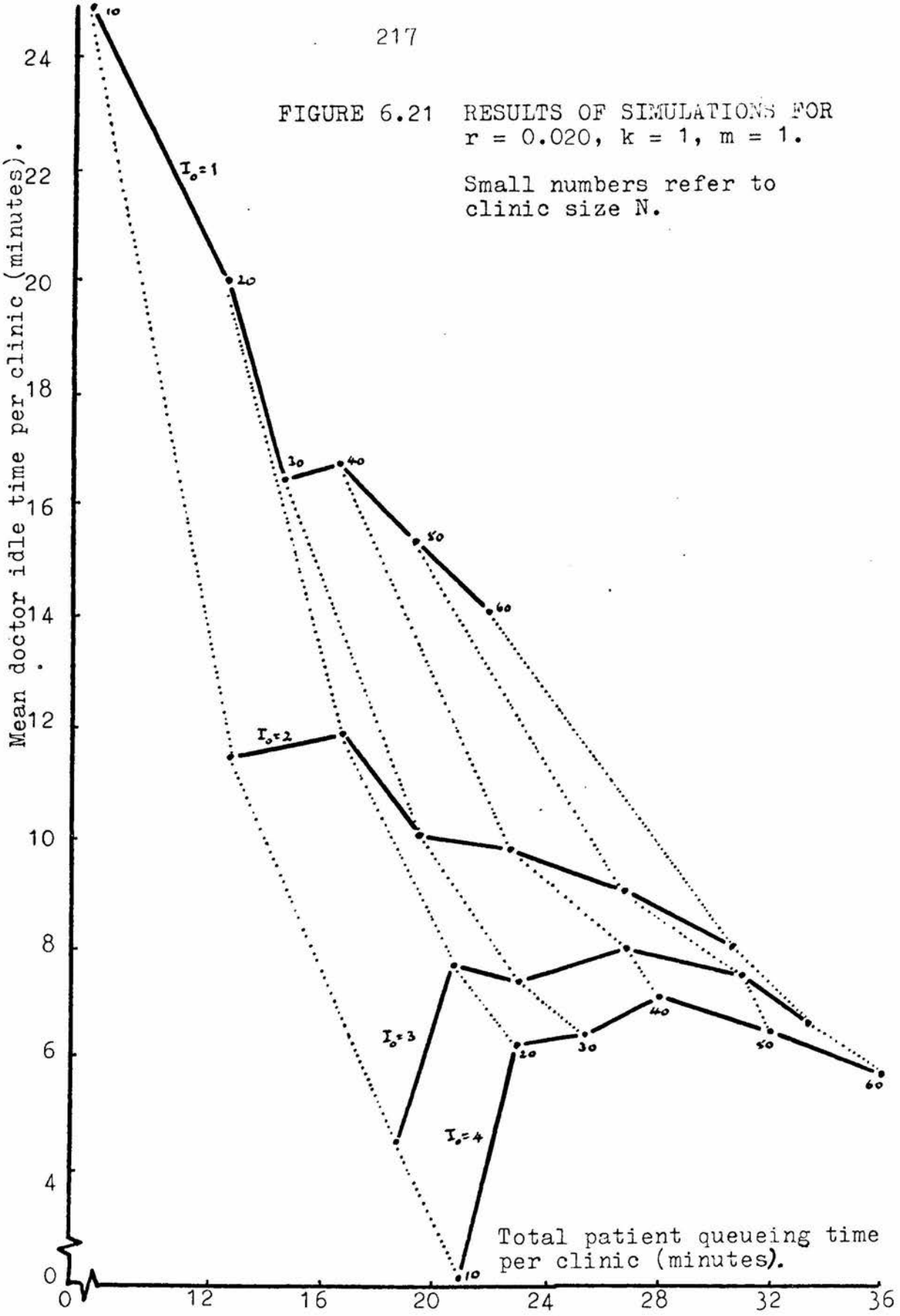


FIGURE 6.20 RESULTS OF SIMULATIONS FOR $r = 0.020$, $k = 2$, $m = 1$.

Small numbers refer to clinic size N.

FIGURE 6.21 RESULTS OF SIMULATIONS FOR $r = 0.020$, $k = 1$, $m = 1$.

Small numbers refer to clinic size N.



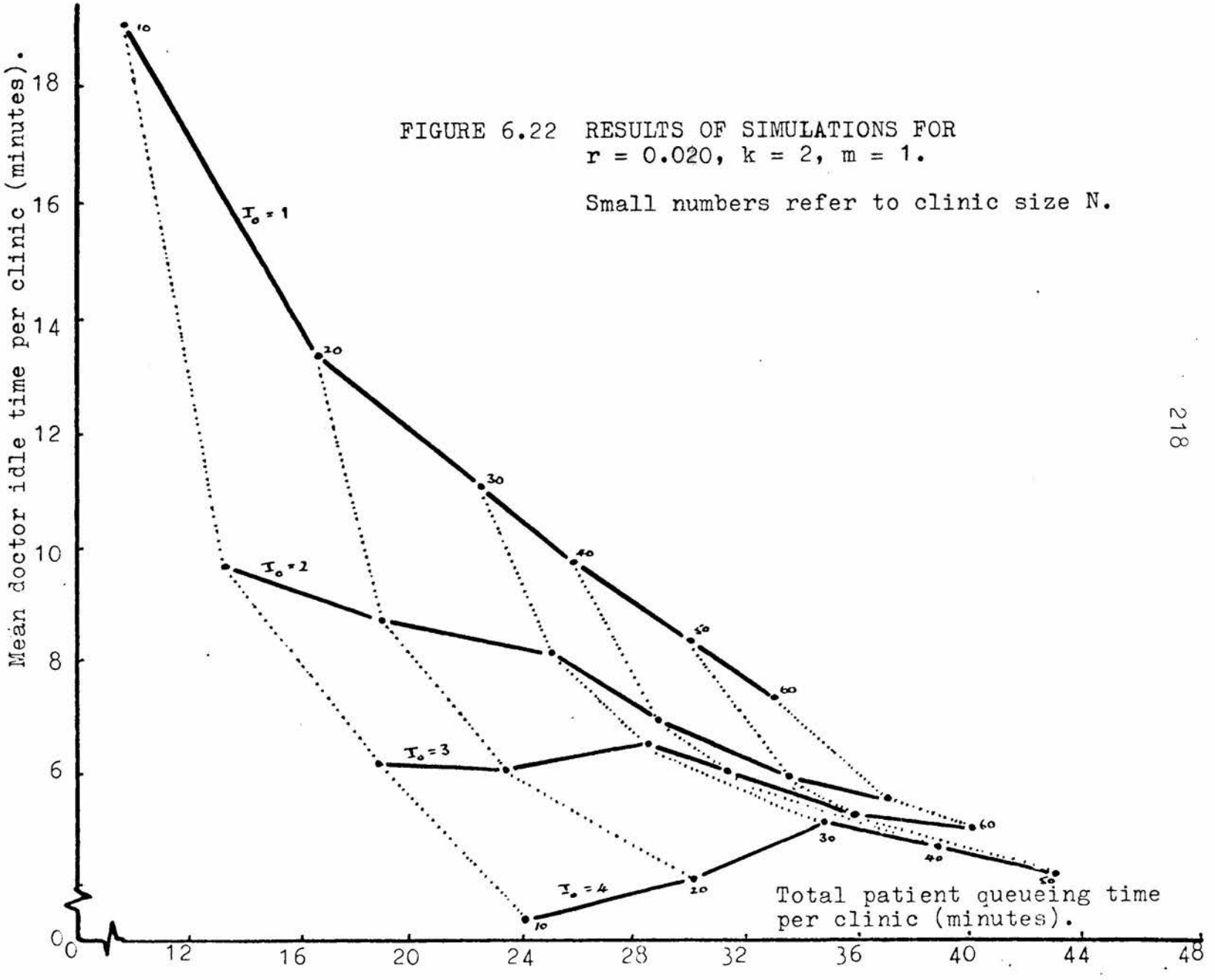


FIGURE 6.22 RESULTS OF SIMULATIONS FOR
$r = 0.020$, $k = 2$, $m = 1$.

Small numbers refer to clinic size N.

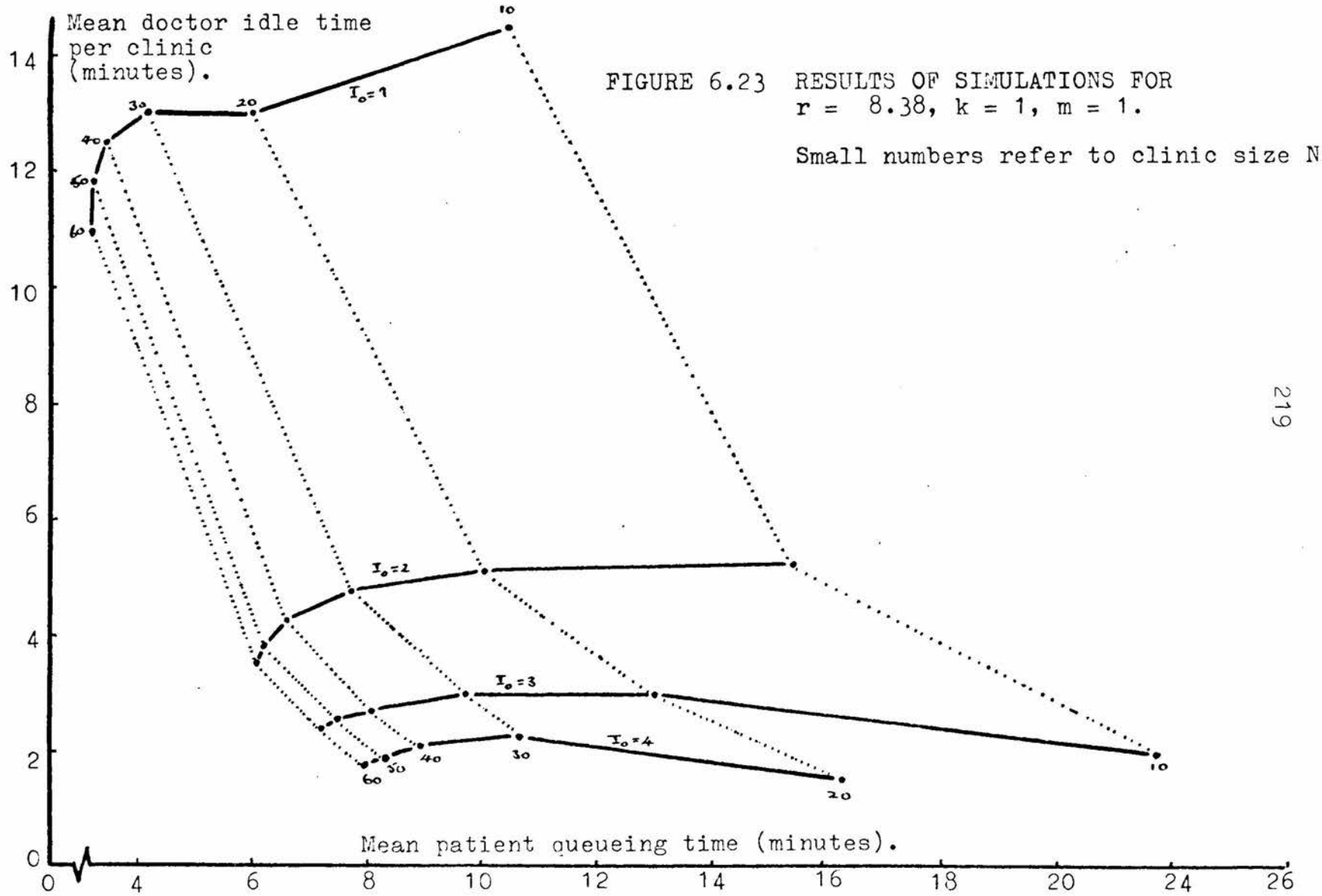
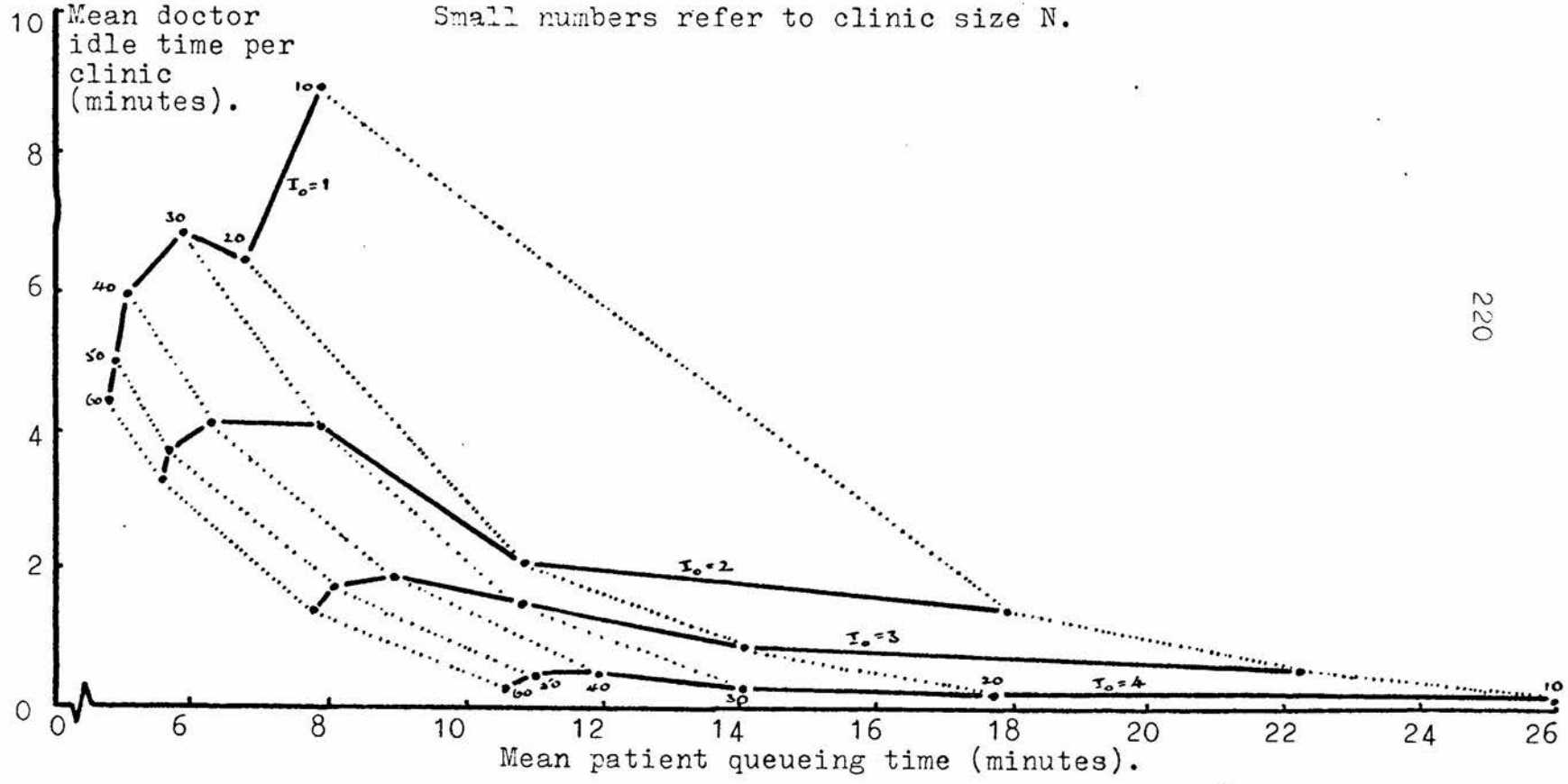


FIGURE 6.23 RESULTS OF SIMULATIONS FOR
$r = 8.38$, $k = 1$, $m = 1$.

Small numbers refer to clinic size N.

FIGURE 6.24 RESULTS OF SIMULATIONS FOR r = 8.38, k = 2, m = 1.



clinics with a smaller initial number, the queueing time is less for the inpatient sessions.

If, in fact, a different value of k applies to each of the inpatient and outpatient groups, then we have a further reason for dealing with the work in segregated sessions. Even assuming the same value for both groups, this section's work has demonstrated that there is a sufficient degree of inhomogeneity between the groups, in the form of different age distributions, to yield substantial positive advantages to work segregation. In practice complete segregation in this manner will probably not be possible; further work might investigate compromise work schedules, for example having sessions dealing for the first half with "inpatients", and the second half "outpatients". Even this will give reduced total idle times and queueing times, on account of the reduced variability of the service time within each half of the session. Research workers in particular hospitals will have to analyse the constitution of the work for all types of examination, but it may be possible to combine the ideas presented here with a simulation program to monitor the efficiency of the appointment schedule in use.

7. Summary and Discussion of Results

In order to establish some of the principles on which the working efficiency of an X-Ray department depends, this work has used a number of descriptive models of both a mathematical and a simulation kind. When interpreting the results, it is important to remember that the real department involves the actions of human beings, rather than inanimate entities; in his particular relationship with the system, each individual will usually seek to gain the greatest benefit to himself. Thus the departmental staff will try to develop a working routine which enables them to carry out their tasks with the least strain on themselves; the patients will try to minimise their time spent at the hospital by complying readily with directions from the staff, and, if possible, by arriving at a slack time of day.

Such actions will tend to reduce, for example, the mean doctor idle-times and patient queueing times from the values indicated by models which do not allow for the intelligence of the individuals concerned. Further, there will be a tendency for staff and patients alike to alleviate the congestion at certain times, and to spread the work load more evenly. This produces a behaviour in the queueing system which is closer to equilibrium than in the equivalent model. Lee (1968) has described this phenomenon thus:-

> "The fact that so many queueing processes involving humans do, quite obviously, appear to be in statistical equilibrium when one knows that the mean input rate is not at all constant, and the service-time distribution is always changing shape, is due to the presence of "beneficient ghosts" which, being effects not allowed for in most mathematical models, yet tend to preserve stability. An example of such a ghost would be an inverse relationship between arrival rate and service time."

Lee's example is but one of the many "ghosts" to be seen to be effective in hospital departments, but it should of course be pointed out that not all "ghosts" are beneficient! However, this digression does illustrate the fact that theoretical models will at best indicate only general principles affecting the system, and not too much emphasis should be placed in the numerical values of their results.

Despite these remarks, the survey work described in Chapter 4 showed that in the Royal Infirmary the main periods of congestion were due at least in part to an uneven pattern of patient arrivals throughout the day. An overload of work was caused at certain hours of the morning and afternoon by the coincidental arrivals of inpatients from the hospital wards and outpatients from specialist clinics. It would appear that any attempts to deal with ward patients at off-peak times or to encourage clinics to send cases at less congested times would be desirable. A particular example of this would be to stagger staff lunch breaks and encourage work to be sent at this time; this would probably ease the congestion of the main mid-morning and mid-afternoon "peaks".

In Section 4.6.3 is described the investigation into the relationship between the time to complete an X-Ray examination, and the age, sex, origin, and mobility of the patient. For various reasons the chest X-Ray was chosen for the main study. The first result to emerge from the data was that old patients had a substantially longer service time than young ones; it was found that the mean service time for patients over 60 was 50% more than that for patients under 60. Differences in this statistic were observed between inpatients and outpatients, and the mean time was also longer for patients

of lesser mobility arriving in wheelchairs and on trolleys. The examination took less time when performed on a miniature film machine. No significant differences were found between the service time means or variances between the various patient groups of opposite sex.

In general in a queueing system, it is desirable to deal with incoming units which have approximately the same service time distribution; if there are some units having a service time distribution with markedly different characteristics, it is preferable to deal with them separately when allowance can be made for these peculiarities. In particular in the hospital clinic it is preferable to deal separately with patients having a longer service time mean; if this is done, an appointment schedule may be adopted having different appointment intervals for the various groups of patients. As we have seen, the service time varies with the age of the patient, being much longer for older people. For practical reasons it is not possible to segregate patients into age groups, nor into classes of mobility. However it is possible to envisage a policy of segregation according to patient origin; by noting the very marked differences in the age distributions of patients from various sources, we may clearly infer that differences in the service time distribution also exist. In Chapter 5 and 6 this idea is developed quantitatively.

The work of the Casualty department does not fit easily into the general pattern of the other work. As discussed below, it would be possible to modify the mathematical and simulation results to consider the relative advantages of having a separate Casualty department as opposed to an integrated department. At the end of Chapter 4, some comments are made on particular

problems observed in the Royal Infirmary.

Chapter 5 concerns mathematical queueing theory models of a hospital clinic, first with a survey of a number of available descriptions, and then with a development of a more general model. Most clinics deal with patients who arrive either mainly by appointment or mainly without appointments. To reflect this, mathematical models often use an input mechanism either of regular arrivals (denoted by D) or random arrivals (M).

A gamma distribution was fitted to the data on service times observed in the Royal Infirmary. This followed the practice of earlier workers in this field. For mathematical simplicity, only integer values of the variance parameter were used, giving the family of Erlangian distributions E_k. Thus the most general models which can be considered so far are $M/E_k/1$ (to represent a clinic with random patient arrivals) and $D/E_k/1$ (regular appointment arrivals). If $k = 1$, the exponential service distribution is obtained.

As a first step, it was decided to consider the queueing model $E_{k_1}/E_{k_2}/1$ because particular values of k_1 and k_2 give the two models described above: if $k_1 = 1$ we have the input process M, and if $k_1 \longrightarrow \infty$ whilst the service time mean is kept constant, we have the input process D. A method of Smith was applied to obtain the Laplace transform of the queueing time distribution; the transform is a function of the complex roots having negative real components of a polynomial equation. In practice these roots would probably have to be found by some iterative or search technique.

The model $E_{k_1}/E_{k_2}/1$ was extended to include technical failures; it is

assumed that the probability of a successful exposure at the j th stage is c_j. For general values of c_j, the expected waiting time is derived by the use of Pollaczek's formula. When it is assumed that the probability of success at any stage is a constant, the polynomial equation takes on a simpler form, and by using partial fractions, the queueing time may be expressed as a weighted sum of exponential variables.

The expected waiting time in the system $M/E_k/1$ is quoted, a result due to Saaty. The waiting time distribution for $E_k/M/1$ is given, and also for $D/M/1$ involving the root of a transcendental equation. The results for both these models are from work by Jackson and Nickols. The more general model $D/E_k/1$ was considered by letting $k_1 \to \infty$ in the system $E_{k_1}/E_{k_2}/1$. The polynomial equation of the latter model becomes a transcendental one, and although simpler in form, iterative methods would again be required in practice to obtain numerical results.

All the above models have an input of units arriving either deterministically at regular intervals, or at random. In an X-Ray department there is usually a mixture of patients arriving with or without appointments, the two classes being represented in some ratio r. It was decided to consider models to allow for this, and an arrival mechanism denoted by $(M + D_m)$ was adopted; this consisted of a stream of arrivals of a mixture of single units at random, and batches of m units at regular intervals. We are thus now considering models of the type $(M + D_m)/E_k/1$.

By selecting at random one of the arrival times and considering the implications of the patient being in either of the two classes, the

distribution of the lengths of the inter-arrival intervals is derived. However the main mathematical difficulty in the solution of this model lies in the fact that successive interval lengths are not independent. The first serial correlation coefficient was considered in the ordered sequence of interval lengths. When all the intervals are used, the coefficient is negative for low values of r, becoming positive for values of r above about 1.0, and converging rather slowly to +1 for very large values of r. It is possible, however, to choose a simple subsequence of the intervals which has a negative correlation always, converging to -1 rapidly with increasing r. For this reason, it was felt that an assumption of independence, for the purposes of mathematical simplicity in the model solution, would be too strong.

The algebraic steady-state solution is derived for the distribution of the number of people in the system just after an appointment time in the model $(M + D_m)/M/1$. This was achieved by noting that during the times between the regular batch arrivals there are fragmentary reaslisations of the simple queue M/M/1. The solution involves an infinite set of equations, and most of the latter half of Chapter 5 describes the approximation and solution of these equations to obtain numerical results.

Each of the coefficients in the above set of equations is a time dependent state-transition probability from the system M/M/1; these probabilities involve an infinite weighted sum of Modified Bessel functions of integer order, and the first problem in the numerical solution of the equations was to find a suitable approximation to this sum. By expanding each Bessel function as a series, it was possible to obtain an upper bound

for the sum; the bound was then used to obtain the finite number of terms of the sum needed for a given order of accuracy in the approximation to the whole.

The second main problem in this numerical solution is in determining how many terms of the distribution $\underline{\pi}$ are to be included in the finite subset of equations to be solved. The criteria adopted were firstly that terms constituting all but a small specified proportion of the probability mass of $\underline{\pi}$ should be included, and secondly that the relative error in the solution for each individual term should be less than some specified amount. The first condition was met by noting that the system $(M + D_m)/M/1$ gives a distribution $\underline{\pi}$ which has both a smaller mean and variance than in the simple queue M/M/1 of the same intensity. The simple queue has a known solution for $\underline{\pi}$ (a geometric distribution), and so it was easy to determine the number of terms with a specified proportion of the probability mass; the same number of terms contains not less than the same proportion in the system $(M + D_m)/M/1$. To consider the relative error of individual terms, the system M/M/1 was solved by the same numerical method as for $(M + D_m)/M/1$; in fact, the simple queue is just the particular model obtained when $r = 0$. By arguments similar to those above, a satisfactory solution to this model was shown to give a satisfactory solution to the general model if the same number of terms are used.

The numerical solution for $\underline{\pi}$ may be used to consider the queueing or waiting time distributions of the patients; in the hospital context it is the queueing time which is more important. For regular arrivals the moments of this distribution are derived in terms of the moments of $\underline{\pi}$.

The quantiles (or percentiles) involve a weighted sum of incomplete gamma functions; by using the fact that the gamma functions are of integer order, the quantiles may be expressed in terms of partial sums of the exponential function. In a practical situation when planning appointment schedules, rather than wishing to know quantiles of the queueing time distribution, we wish to determine the probability that a patient will queue longer than a certain specified time; thus it is a probability which is found, rather than a time. These are the results presented in section 5.8.

The properties of the queueing time distributions for the random arrivals are much harder to derive, partly because the distribution applies to instants just after regular arrivals, and partly because of the difficult mathematical form of the simple queue transition probabilities. It was possible to show that the expected queueing time for random arrivals is not greater than the expected waiting time for regular arrivals, and that it is equal to the expected queueing time in an equivalent simple queue. An expression is given for the quantiles of the distribution, which are also functions of the time since the last regular arrival. The form of the variance of this distribution is also shown.

The time-dependent solution is derived for the model $(M + D_m)/M/1$; the solution is in the form of a series of distributions of the numbers in the system at the instants after the first few regular arrivals. To see how rapidly the queueing process seemed to settle down to an equilibrium behaviour, the first two moments of each distribution were compared to the corresponding values from the equilibrium solution developed earlier. Ideally, when comparing two distributions in this way, we would like to have

values for all the moments of each. In this case, however, we are making a comparison of two approximations to some unknown distribution, and it is clearly unreasonable to expect even a large finite number of moments to be in close agreement. We are concerned here with an approximate measure of the rate of convergence, and the first two moments seem to serve the purpose.

In the numerical results it was found that more rapid convergence occurs with large values of the batch size m, and low values of r and the traffic intensity ρ . This is obviously to be expected for ρ , and it is thought that the effect of m is reasonable. However there seems to be no intuitive reason why high values of r should give less stability in the process. Other results of this chapter showed that high values of r gave shorter queueing times on average. It might be possible, in clinics of this type, to call patients at a faster rate at the start of the session, thereby still attaining equilibrium within a fairly short time, without increasing the queueing times beyond reasonable limits. Also, if the opening remarks of this chapter are borne in mind, there is reason to suppose that an equilibrium condition may be reached more rapidly in the real department than indicated by Table 5.14, even allowing for extreme circumstances or a changing service time distribution.

Theoretical models of the kind described above have limited usefulness when planning an appointment schedule in practice. This is not to say they are totally without value, because they do provide a theoretical foundation for the principles on which more practical studies may be based. In a real clinic, however, there will always be empirical variation, which may be a function of the particular clinic, and which will almost certainly deny a

theoretical analysis. We are thus driven to use methods of a more empirical nature, such as the simulations described in Chapter 6.

The first group of results extend the work of Chapter 5 to include clinics where the traffic intensity is one. Such systems cannot exhibit stable equilibrium behaviour, and so finite realisations of the queueing process are considered. As the parameters r and ρ apply to "infinite" steady-state conditions, biases appear in their observed values in such finite experiments. The nature of the biases depend on the simulation stopping rule; it was decided here to compare clinics of a constant size, and the particular biases which result are considered in section 6.4. It is shown that the bias in r is only negative if the batch size m is greater than the initial number I_o; this is unlikely, because, in a rational appointment schedule the rate of calling patients will be faster at the start of a session than at other times, and so I_o will be greater than m. It is also shown that an upper bound for the bias in r is given (in the notation of section 4.6) by:-

$$\varepsilon_1 \quad < \quad I_o r/mN^* \ .$$

N^* will always be of the same order as N: thus for a small clinic, taking $I_o = 2$, $N = 10$, $m = 1$, the worst relative bias in r will be 20%; for a clinic with $I_o = 3$, $N = 60$, $m = 1$, this will be 5%. We should thus bear in mind that the results for small clinics will tend to favour the doctor, having rather lower idle time than a clinic dealing with patients in the two classes in the "true" ratio r. The differences for large clinics will all be small, but we may tend to slightly underestimate the mean patient queueing time.

This last difficulty with small clinics is simply part of the general picture when the results of section 6.7 are considered. For most values of the defining parameters, it was only some clinics of size 10 and 20 which did not meet the Ministry of Health's suggested standard; these were usually clinics with a high value of I_o. On the other hand, a low initial number I_o corresponded to a rather high value for the doctor idle-time. In short, such clinics do present the greatest problems in the adoption of an appointment schedule, and great care must be taken when choosing the value of I_o to be used. It is ironic that it is clinics with a small number of patients, and thus probably long examinations, which present the most difficulty; it is precisely these clinics for which appointment schedules are most desirable, as each patient represents more staff effort than in a clinic with a shorter mean examination time.

There seems to be no intuitive explanation of the "hump" shape of response surface to be seen in higher values of I_o (as in Figure 6.2), but as it is present in many sets of independent simulations, it does seem that this is representative of the real surface, and is not to be explained by sampling errors.

Among the general remarks made on the results of Figures 6.2 to 6.6 is that with low values of I_o, it is mainly the doctor idle-time that varies with the clinic size, but for higher values of I_o, the patient queueing time is more affected. Figures 6.7 to 6.9 show results for different values of r. It was observed that there are nearly always substantial reductions in either the doctor idle-time or the mean patient queueing time, or both, when r is increased; for small clinics it is usually the former that shows the

greater decrease, and the latter for large clinics. The effect of the batch size m is considered in Figures 6.10 to 6.12; the differences in idle times and queueing times for different values of m are small for large clinics, but in general it is difficult to justify using values of m greater than one other than in a few small clinics having a very variable service time distribution. At the end of section 6.7, there is a demonstration of how to use the simulation results to derive optimal values of I_o, when the other parameters of the clinic are given, and a quantification made of the relative values of the times wasted by a doctor and his patients.

The second group of simulations, described in section 6.8, includes a number of additional refinements concerned with a more detailed analysis of the variation of service times. The earlier work assumed the same service time distribution for all patients, whereas here the survey work of Chapter 4 is used to specify a service time which is a function of the patient's origin, age and mobility. The input to the simulated clinics is taken as being mainly of one type, either inpatient or outpatient; the treatment given is somewhat idealised, but in almost any hospital, it would be possible to name, even without a detailed survey, groups of patients which are known from experience to have service time distributions which are substantially different in character from the distribution applying to the rest of the patient input. If such cases can be scheduled separately, efficiency will always increase.

The simulations discussed here involve dealing with patients in sessions of two main types, either inpatient or outpatient. This is a very simple grouping of the work, and probably easy to implement in practice. In

any particular clinic, it would be possible to make a more detailed analysis of the work load for each examination or treatment, and group the patients accordingly. Also accurate estimates could be made of the proportions of each class of patient having appointments, which here have assumed values.

The results, plotted in Figures 6.19 to 6.24 show the same general patterns as in the earlier work. We may see that usually the inpatient session is more efficient than the corresponding outpatient session; hence in this example the longer service time mean of the inpatients has been well compensated by the much higher value of r which is possible in the inpatient clinic.

7.1 Conclusion

In the field of mathematical queueing theory, the investigation of systems with more than one input source is very new. Consideration has been given to models where it is only the arrival mechanism that differs between the inputs, and once the material or units are in the system, there is no need to differentiate between them. An example of this is a dam or storage situation, such as considered by Sahin (1971). However in the hospital clinic, there is a need to distinguish between the waiting, queueing, and service time distributions of the two general classes of patient, i.e. appointment and random, and future theoretical work in this area should allow for this distinction.

In the model $(M + D_m)/M/1$, the derivation of a closed form expression for the terms of the distribution $\underline{\pi}$ appears to be made very difficult by the rather complicated mathematical structure of the simple queue time-

dependent state-transition probabilities. However, a simpler problem might be that of obtaining closed expressions for the moments of the queueing time distribution for regular arrivals, and for the variance of the same distribution for random arrivals. The numerical results of Tables 5.7, 5.9 and 5.10 might be extended by using equation (5.37) to obtain the equivalent results for random arrivals.

In the simulation work, a "first in, first out" queue discipline was used. For clinics dealing with emergency casualty work, this is clearly inadequate and further simulations would be needed in which such patients were given a high priority, possibly pre-emptive. In such a situation, the distinction between the queueing time distributions of the various classes of patient becomes important, and it would be necessary to specify acceptable levels of queueing for both patient groups before an appointment schedule could be chosen. It would be possible by such means to decide what was the minimum proportion of casualty work which would make the provision of separate emergency facilities desirable; by simulation one could compare the characteristics of the equivalent separated and integrated departments. A similar distinction in the patient classes might also be needed where it was clear that a certain type of patient needed to be dealt with more urgently than the others.

Further refinements to the simulation model of Chapter 6 could be made by fitting accurately a different service time distribution for each age group of each of the various patient groups. Time was not available to collect sufficient data to do this here, and a constant value of the service time variance parameter k was assumed for the whole of a given clinic session.

It may, of course, be unreasonable to fit Erlangian distributions for each class, and it may be necessary to use empirical distributions instead. This is not important for simulation purposes, as there will probably be no significant difference in the computer time needed to select an observation from an empirical distribution rather than generate one from a convenient theoretical distribution. However, the inclusion of mixed service time distributions, even if all of one type (such as Erlangian), creates considerable difficulties in a theoretical model; in particular a model with a general mixed service distribution would require a simulation analysis at the present time. In practice, simulation models may also include some local details, possibly being used periodically to monitor the effectiveness of appointment schedules. The methods of Jeans et al (1972) and the ideas presented here might be extended to predict the effects of major changes in a department's working environment, an example being the introduction of a telecommunications system to increase room occupancies by using a central receptionist to direct work to a suitable area of the department.

With the present state of knowledge, it would appear that simulation and other empirical techniques can offer a much more rapid answer to problems presented in clinic administration. However the theoretical side of the work is also important, and if a solution to the general many server model is found, the theoreticians will have provided a further basis for the development of existing practical techniques in this area.

Appendix 1

"Towards a Clearer View : The Organisation of Diagnostic X-Ray Departments."

Published for the Nuffield Provincial Hospitals Trust by the Oxford University Press, 1962.

(Extract)

Principles Evolved from the Study and Summary of Recommendations to Implement Them

The diagnostic X-Ray department's organization should seek to satisfy demand by the most effective and economical use of its human and material resources. It would appear from this study that the practice in hospitals falls short of this ideal. Hospital authorities are therefore strongly recommended to examine the organization of their diagnostic X-Ray departments, and to apply such of the following recommendations as their findings show to be appropriate. They are especially urged to give effect to these principles in planning new departments.

Principles

(1) As much as possible of the work coming to the department should be controlled by a well conceived appointments system.

(2) There should be close co-operation in the organization of work between the X-Ray department and the other departments in the hospital.

(3) Control of the work-load within the department should be centralized.

(4) Full advantage should be taken of the versatility of X-Ray apparatus.

(5) Radiographers should be employed on duties which demand their specialized skill.

(6) As far as possible the X-Ray facilities of a hospital should be centralized in one carefully sited department.

(7) X-Ray departments should be designed and equipped to save labour.

(8) The staff for emergency services should be drawn from as many hospitals as practicable.

(9) There should be regard for the patients' point of view.

Summary of Recommendations to Hospital Authorities

(1) The work-load on the X-Ray department should be assessed and analysed so that information is available about the sources and incidence of the department's demands.

(2) As much of this work as possible should be made subject to the control of an appointments system.

(3) Some indication of priority should be put on inpatients' X-Ray requests.

(4) The work-load emanating from outpatient clinics should be studied in consultation with the staff of the clinics themselves so that the incidence of the demands from this source may be predicted.

(5) This prediction having been made, the appointments system (2 above) should be used to regulate the work-flow to fit the availability of the department's resources.

(6) There should be a suitably trained and able receptionist to receive and distribute the work-load under the direction of the Superintendent Radiographer and the Director of the Department.

(7) The versatility of X-Ray apparatus should be exploited to the full and, where possible, increased by fitting ancillary equipment.

(8) There should be flexibility in the use of rooms so that a patient may be examined on any suitable set of apparatus which is available. (NOTE: This does not preclude the allocation of blocks of work to

X-Ray rooms, or the planning of their day in sessions.)

(9) The use made of radiographers' skill and time should be examined and they should be relieved of those tasks which can be carried out by ancillary help.

(10) The system of checking radiographs should recognise the importance of encouraging the radiographers' interest and sense of responsibility. They should be responsible for the marking of X-Ray films.

(11) Labour saving devices should be installed to increase the efficiency with which radiographic and clerical work is done. New departments should be designed to reduce the distances staff have to cover during their work.

(12) The present arrangements for dealing with the request, the film and the report should be reviewed to take account of the recommendations of the Adrian Committee and the higher rate of output which will result from better organization and automatic film processing.

(13) Modern business methods should be applied to the filing and distribution of films and to the other clerical work of the department.

(14) The arrangements for providing emergency services should be reviewed with the object of sharing its demands among as many radiographers as possible, perhaps on an area basis.

(15) In each department where the director delegates administrative responsibilities to the Superintendent Radiographer such delegated responsibilities should be clearly defined. There should also be a training course to help Superintendent Radiographers in the discharge of their administrative duties.

(16) Experiments should be conducted to test the possibilities of extending the length of the department's working day.

(17) Occupancy of waiting space by inpatients should be kept at a minimum.

(18) Patients should be fully informed before and during their visits to X-Ray departments why they are there, what will happen to them and, if need be, why and for how long they have to wait.

Appendix 2

"Waiting in Outpatient Departments".

Published for the Nuffield Provincial Hospitals Trust
by the Oxford University Press, 1965.

(Extract)

Summary of conclusions and recommendations

Timings were taken of approximately 12,500 outpatients and 900 doctors at 474 clinics in various specialities at 60 hospitals.

The average time waited by patients after appointment time before seeing a doctor was 25 minutes; 11% of the patients had to wait for over an hour and 34% for over half an hour. At only 11 of the 60 hospitals was waiting time within the limits suggested by the Ministry of Health.

The time which doctors had to wait during clinics because patients were late or failed to arrive was negligible - on average less than one minute per doctor.

Almost 13% of booked patients failed to keep their appointments without giving adequate notice of doing so. On the other hand, 4% of patients seen did not have appointments.

On average, patients arrived at the outpatient department six minutes earlier than their appointment times; almost a third of the patients arrived late but only one in every ten was more than a quarter of an hour late.

Outpatients travelling by ambulance tended to arrive later than the remainder of patients and their individual unpunctuality was more widely dispersed.

The failure of doctors to attend their clinics was negligible.

Only one clinic in every five started early or on time and on average clinics were 12 minutes late in starting. At only one hospital out of the 60 did all the clinics observed start punctually on average.

Doctors were generally late starting to see patients - by 15 minutes on average. Thirty-nine per cent of doctors were more than 15 minutes late and 14% more than half an hour late.

On the whole the number of patients per doctor booked for the start of the clinic was not excessive, but the advantage of this was usually lost because clinics started late.

At only a minority of the hospitals were block bookings avoided or kept reasonably low. The appointment systems of 25% of all the clinics observed were badly designed in allowing excessively large block bookings and a further 50% of the systems were capable of improvement in this respect.

In over half the clinics patients' appointment times were condensed into too short a period with the result that the patients arrived more quickly than the doctors could deal with them.

Appointment systems cannot be properly designed without mean consultation times first being measured. Several methods of doing this are described. When mean consultation time is known, the length or size of a clinic can be fairly accurately determined.

Not more than three patients per doctor should normally be booked for the start of a clinic.

Individual appointment systems are recommended, but block bookings are sometimes justified in clinics where consultation times are very short - when this is so the blocks should be not larger than three patients per doctor.

The lateness of doctors and the poor design of appointment systems appeared to be the two major causes of patients' waiting.

Lateness on the part of medical staff poses an administrative problem and the remedy depends partly on the skill of the hospital administrator and partly on the willingness of the doctor to change his habits.

Waiting time can be measured and reduced by quite simple methods, particularly in clinics where the appointment systems are grossly inaccurate.

Maintenance of appointment systems should be a matter of routine administration and should be carried out regularly.

The problems presented by variable consultation times, teaching clinics, preliminary tests and transport difficulties are not insuperable when designing appointment systems.

Patients should be seen, as far as possible, in appointment order but late patients should not always be penalised by being made to wait until the end of a clinic.

It is not accepted that the use of appointment systems causes waiting lists.

Because small hospitals did not appear to have problems in respect of their appointment systems, they were excluded from the survey.

Waiting times were no shorter in new outpatient departments than in old ones; it was noted that new departments were not without faults in design, sometimes quite serious.

Bibliography

Ashworth, W. J. (1954). Fatigue reduction in the X-Ray department. Radiography, 20, 142.

Bailey, N. T. J. (1952a). A study of queues and appointment systems in hospital outpatient departments, with special reference to waiting times. J. R. Statist. Soc., (B), 14, 185.

Bailey, N. T. J. (1952b). Operational Research in medicine, O.R. Quarterly, 3, 24.

Bailey, N. T. J. (1954a). On queueing processes with bulk service. J. R. Statist. Soc., (B), 16, 80.

Bailey, N. T. J. (1954b). Queueing for medical care. Applied Statistics, 3, 137.

Bailey, N. T. J. (1955). A note on equalising the mean waiting times of successive customers in a finite queue. J. R. Statist. Soc., (B), 17, 262.

Bailey, N. T. J. (1956). Statistics in hospital planning and design. Applied Statistics, 5, 146.

Bailey, N. T. J. (1957). Operation Research in hospital planning and design. O.R. Quarterly, 8, 3, 149.

Blanco White, M. J. and Pike, M. C. (1964). Appointment systems in outpatient clinics and the effect of patients' unpunctuality. Medical Care, 2, 133.

Cox, D. R. and Smith, W. L. (1954). On the superposition of renewal processes. Biometrika, 41, 91.

Cox, D. R. and Smith, W. L. (1967). "Queues". Methuen, London.

Fetter, R. B. and Thompson, J. D. (1965). The simulation of hospital systems. Op. Research, 13, 689.

Fisher, R. A. and Yates, F. (1957). Statistical tables for biological, agricultural and medical research. (Fifth edition). Oliver and Boyd, Edinburgh and London.

Flagle, C. D. (1959). The problem of organisation for hospital inpatient care. Paper given at the 6th international meeting of the Institute of Management Sciences, Paris, 1959. Pergamon Press.

Fraser, B. J. (1969). The organisation of a radiology department in a district general hospital. Ph.D thesis, University of Reading.

Gabrielson, I. W., Sorians, A., Taylor, M.M. and Flagle, C. D. (1959). Analysis of congestion in an outpatient clinic. O.R. Division, John Hopkins Hospital, Baltimore, 5, Maryland, U.S.A.

Gaver, D. P. (1959). Imbedded Markov chain analysis of a waiting line process in continuous time. Ann. Math. Statist., 30, 3, 698.

Hardie, M. C. (1955). Waiting by outpatients. Hospital, 51, 763.

H.M.S.O. (1962). A hospital plan for England and Wales, Cmnd. 1604.

H.M.S.O. (1968). Public expenditure 1968-69 to 1973-74. Cmnd 4234.

H.M.S.O. (1970). X-Ray departments : work of radiographers. Abstracts of efficiency studies in the health services, No. 141.

The Hospital (November, 1958). Editorial.

Jackson, R. R. P., and Nickols, D. G. (1956). Some equilibrium results for the queueing process E/M/I. J. R. Statist. Soc., (B), 18, 275.

Jackson, R. R. P., Welch, J. D., and Fry, J. (1964). Appointment systems in hospitals and general practice. O.R. Quarterly, 15, 219.

Jeans, W. D., Berger, S. R., and Gill, R. (1972). Computer simulation of an X-Ray department. Brit. Med. Jour., 1, 675.

Kendall, D. G. (1951). Some problem in the theory of queues. J. R. Statist. Soc., (B), 13, 151.

Kendall, D. G. (1953). Stochastic processes occuring in the theory of queues and their analysis by means of the imbedded Markov chain. Ann. Math. Statist., 24, 338.

Kingman, J. F. C. (1966). The algebra of queues. J. App. Prob., 3, 285.

Lee, A. M. (1968). "Applied Queueing Theory". Macmillan.

Lindley, D. V. (1952). The theory of queues with a single server. Proc. Camb. Phil. Soc., 48, 277.

Luck, J. H. (1955). Outpatient waiting time. Medical Record, 3, 430.

Ministry of Health (1958). Outpatient waiting times. Hospital O. and M. reports, No. 1.

Nuffield Provincial Hospitals Trust (1955). "Studies in the function and design of hospitals. "Oxford University Press.

Nuffield Provincial Hospitals Trust (1962). "Towards a Clearer View: the organisation of diagnostic X-Ray departments. "Oxford University Press.

Nuffield Provincial Hospitals Trust (1965). "Waiting in outpatient departments : a survey of outpatient appointment systems." Oxford University Press.

The Office of Health Economics (1963). Hospital costs in perspective. Pamphlet No. 3.

The Office of Health economics (1967). Efficiency in the hospital service. Pamphlet No. 22.

The Office of Health Economics (1970). Building for health. Pamphlet No. 35.

Olver, F. W. J. (1962). Tables for Bessel functions of moderate or large orders. National Physical Laboratory Tables, Vol. 6, H.M.S.O., London.

Pearson, K. (1965). Tables of the Incomplete Γ - Function. Biometrika Tables, Cambridge University Press.

Pike, M. C. (1963a). Waiting in hospital outpatient and casualty departments. Ph.D. thesis, University of Aberdeen.

Pike, M. C. (1963b). Some numerical results for the queueing system $D/E_k/1$. J. R. Statist. Soc., (B), 25, 477.

Roberts, P. A. (1956). The planning of X-Ray departments in relation to other departments in a hospital. Radiography, 22, 78.

Rossiter, C. E., and Reynolds, J. A. (1963). Automatic monitoring of the time waited in outpatient departments. Medical Care, 1, 4, 218.

Saaty, T. L. (1961). "Elements of Queueing Theory". McGraw Hill, New York.

Sahin, I. (1971). Equilibrium behaviour of a stochastic system with secondary input. J. App. Prob., 8, 2, 252.

Sahin, I., and Bhat, U. N. (1971). A stochastic system with scheduled secondary inputs. Op. Research, 19, 2, 436.

Scott, R., and Gilmore, M. (1966). The Edinburgh hospitals. Problems and Progress in Medical Care. Nuffield Provincial Hospitals Trust; Oxford University Press.

Smith, J. (1968). "Computer Simulation Models." Griffin.

Smith, W. L. (1953). On the distribution of queueing times. Proc. Camb. Phil. Soc., 49, 449.

Tocher, K. D. (1963). "The Art of Simulation." English Universities Press.

U.S. Department of Commerce (1967). National Bureau of Standards : "Handbook of Mathematical Functions." Applied Maths. Series, No. 55. Chapter 9, "Bessel Functions of Integer Order".

Welch, J. D. (1952a). Some research into the organisation and design of hospital outpatient departments. J.R. San. Inst., 72, 4, 298.

Welch, J. D. (1952b). Marshalling and queueing hospital applications. O.R. Quarterly, 3, 8.

Welch, J. D., and Bailey, N. T. J. (1952). Appointment systems in hospital outpatient departments. The Lancet, 262, 1105.

Williams, E. R. (1945). The planning of the diagnostic radiological department in a large general or teaching hospital. Brit. Jour. Radiol., 18, 267.